

Fall 1996

Patterns of base composition within and between animal mitochondrial genomes

Nicole T. Perna

University of New Hampshire, Durham

Follow this and additional works at: <https://scholars.unh.edu/dissertation>

Recommended Citation

Perna, Nicole T., "Patterns of base composition within and between animal mitochondrial genomes" (1996). *Doctoral Dissertations*. 1921.

<https://scholars.unh.edu/dissertation/1921>

This Dissertation is brought to you for free and open access by the Student Scholarship at University of New Hampshire Scholars' Repository. It has been accepted for inclusion in Doctoral Dissertations by an authorized administrator of University of New Hampshire Scholars' Repository. For more information, please contact nicole.hentz@unh.edu.

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

**PATTERNS OF BASE COMPOSITION WITHIN AND BETWEEN ANIMAL
MITOCHONDRIAL GENOMES**

BY

**NICOLE T. PERNA
B.S., BIOLOGY, THE PENNSYLVANIA STATE UNIVERSITY, 1991**

DISSERTATION

**Submitted to the University of New Hampshire
in Partial Fulfillment of
the Requirements for the Degree of**

Doctor of Philosophy

in

Genetics

September, 1996

UMI Number: 9703368

**Copyright 1996 by
Perna, Nicole T.**

All rights reserved.

**UMI Microform 9703368
Copyright 1996, by UMI Company. All rights reserved.**

**This microform edition is protected against unauthorized
copying under Title 17, United States Code.**

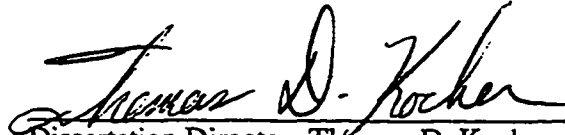
UMI
300 North Zeeb Road
Ann Arbor, MI 48103


ALL RIGHTS RESERVED

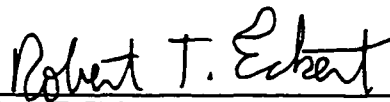
c 1996

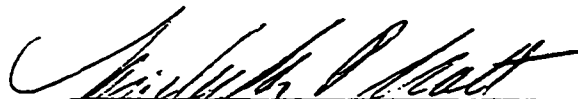
Nicole T. Perna


This dissertation has been examined and approved.


Dissertation Director, Thomas D. Kocher
Associate Professor of Zoology and Graduate
Program in Genetics


Anita S. Klein
Associate Professor of Biochemistry and
Molecular Biology and Chair of Graduate
Program in Genetics


Robert T. Eckert
Associate Professor of Natural Resources
and Graduate Program in Genetics


Michelle P. Scott
Associate Professor of Zoology


Marie A. Gaudard
Professor of Statistics

6/21/91
Date

ACKNOWLEDGEMENTS

I would like to thank my dissertation advisor, Tom Kocher, for his participation, interest, support and often unsolicited good advice. I am also grateful to the rest of my committee, Anita Klein, Marie Gaudard, Michelle Scott and Bob Eckert, for their varied contributions to my completion of this degree and education as a whole. Many other faculty of the Genetics Program, Department of Biochemistry and Department of Zoology have been influential. I greatly appreciate the encouragement of the postdocs, graduate students, and other denizens of the Kocher lab, Scott France, Patty Rosel, Woo-jai Lee, Jeff Markert, Janet Conroy and Karen Carleton, as well as that of other students and friends. I would also like to thank the Graduate School for support throughout the tenure of my stay at the University of New Hampshire. Special thanks to my extended family for unwaivering confidence and to Jeremy, for being Jeremy.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iv
LIST OF TABLES	viii
LIST OF FIGURES	x
ABSTRACT	xii

CHAPTER I

INTRODUCTION TO BASE COMPOSITION, UNEQUAL BASE SUBSTITUTION AND PHYLOGENETIC RECONSTRUCTION	1
Mitochondrial DNA Evolves by Base Substitution	2
Phylogeny Reconstruction, Models of Base Substitution, and Base Composition	4
Beyond phylogenetic reconstruction	14

CHAPTER II

STRAND-SPECIFIC PATTERNS OF NUCLEOTIDE COMPOSITION AT FOURFOLD DEGENERATE SITES OF ANIMAL MITOCHONDRIAL GENOMES	16
Introduction	16
Methods	18
Sequences	18
Codon frequencies	19
Statistics	20
Results and Discussion	21
Nematodes	24
Insects	24
Molluscs	25
Vertebrates	25
Echinoderms	25
Correlations among the statistics	25
Variation within genomes	28
Conclusions	29

CHAPTER III

LOG-LINEAR ANALYSIS OF SYNONYMOUS BASE COMPOSITION: MUTATIONAL BIASES AND TRANSLATIONAL LEVEL SELECTION IN MAMMALIAN MTDNA	30
Introduction	30
Methods	34
Categorical Variables	34

Log-linear Models	35
Results and Discussion	37
Residual Analysis	40
Conditional Independence Models and Evolutionary Hypotheses	42
Base Composition and Position in the Genome	45
Base Composition and Codon Family	48
Higher-order Interactions	51
Dinucleotide and Higher-order Mutational Biases	52
Other sources of Selection on Synonymous Sites	53
Conclusions	54

CHAPTER IV

INTRAMOLECULAR PATTERNS OF SYNONYMOUS BASE COMPOSITION IN TWO ADDITIONAL TAXONOMIC GROUPS: INSECTS AND MOLLUSCS	55
Introduction	55
Methods	60
Results and Discussion	63
<i>Katharina tunicata</i>	63
<i>Cepaea nemoralis</i>	67
Mollusc summary	69
<i>Drosophila yakuba</i>	71
<i>Anopheles gambiae</i>	77
<i>Apis mellifera</i>	81
Insect Summary	86
Conclusions	92

CHAPTER V

STRAND-SPECIFIC DIRECTIONAL MUTATION PRESSURES AND THE COMPOSITION OF MITOCHONDRIAL PROTEINS	93
Introduction	93
Directional Mutation Pressure Theory	94
Strand-Specific Directional Mutation Pressures	95
Mitochondrial Mutation Pressures	96
Methods	97
Results	100
Mutational Pressures at Fourfold Degenerate Sites	100
Mutational Pressures at First and Second Codon Positions	108
Variation in Response to Mutational Pressures among Non- homologous Genes	112
Variation in Response to Mutational Pressures among Amino Acids	118
Discussion and Conclusions	124

LITERATURE CITED	130
APPENDICES	138
Appendix A: PASCAL Source Code for USAGE	138
Appendix B: Sample Datarange Input File	158
Appendix C: Sample Output Log	159

LIST OF TABLES

Table 1.1. Thirteen mitochondrially encoded proteins	4
Table 1.2. Substitution matrix for the human mitochondrial control region	11
Table 2.1. Nucleotide composition of fourfold degenerate third codon positions from 16 animal taxa.	22
Table 3.1. Exact boundaries of the discrete position categories for the variable (D) . . .	35
Table 3.2. Overall composition of fourfold degenerate third codon positions from 12 mitochondrial genes encoded on one strand in the cow, fin whale, harbor seal and human genomes.	37
Table 3.3. Nineteen possible log-linear models	39
Table 3.4. The significance of the [DB] term in log-linear models of the 5 codon family data.	43
Table 3.5. The significance of the [CB] term in log-linear models of the 5 codon family data.	43
Table 3.6. The significance of the [SB] term in log-linear models of the 5 codon family data.	44
Table 3.7. Log-linear models to test the significance of [DB] and [CB] terms for each mammal individually	45
Table 4.1. Summary data for log-linear models fit to the <i>Katharina tunicata</i> data . . .	64
Table 4.2. The significance of two-way interaction terms involving base for log- linear models of <i>Katharina tunicata</i> fourfold degenerate sites.	65
Table 4.3. <i>Katharina tunicata</i> fourfold degenerate site base composition for each codon family on each strand	66
Table 4.4. Summary data for log-linear models fit to the <i>Cepaea nemoralis</i> data	67
Table 4.5. The significance of two-way interaction terms involving base in log- linear models of <i>Cepaea nemoralis</i> fourfold degenerate sites.	68
Table 4.6. <i>Cepaea nemoralis</i> fourfold degenerate site base composition for each codon family	69
Table 4.7. Summary data for log-linear models fit to the <i>Drosophila yakuba</i> data . . .	72

Table 4.8. <i>Drosophila yakuba</i> fourfold degenerate site base composition for each codon family on each strand	76
Table 4.9. Summary data for log-linear models fit to the <i>Anopheles gambiae</i> data . . .	78
Table 4.10. <i>Anopheles gambiae</i> fourfold degenerate site base composition for each codon family on each strand	80
Table 4.11. Summary data for log-linear models fit to the <i>Apis mellifera</i> data	82
Table 4.12. <i>Apis mellifera</i> fourfold degenerate site base composition for each codon family on each strand	85
Table 5.1. Taxa, GenBank accession numbers and citations for 31 complete metazoan mitochondrial genomes.	99
Table 5.2. Summary output from simple linear regressions of the relative frequency of a base at the first or second codon positions regressed on the relative frequency at fourfold degenerate sites	111
Table 5.3. Summary of 40 simple linear regressions of the relative frequency of an amino acid regressed on the relative frequency of a base found in either the first or second position of the codon for that amino acid	117

LIST OF FIGURES

Figure 1.1. Alternative character state reconstructions and substitution asymmetry . . .	13
Figure 2.1. Phylogenetic distribution of %GC ,GC-skew and AT-skew	23
Figure 2.2. Scatterplots of AT-skew vs. %GC, GC-skew vs. %GC, and GC-skew vs. AT-skew	27
Figure 3.1. Boxplot of standardized residuals of the [SCD][SB][CB][DB] model . . .	41
Figure 3.2. Estimated leverages of the [SCD][SB][CB][DB] model	41
Figure 3.3. Predicted percent of the four nucleotides in each distance class for the cow, fin whale, harbor seal, and human fourfold degenerate third codon positions	47
Figure 4.1. Mutation-selection equilibrium frequency of an advantageous character state for two underlying mutational pressures	58
Figure 4.2. Linearized map of the <i>Cepaea nemoralis</i> genome illustrating the boundaries of the four classes of the categorical variable (D)	62
Figure 4.3. Linearized map of the <i>Katharina tunicata</i> genome illustrating the boundaries of the four classes of the categorical variable (D)	62
Figure 4.4. Linearized map of the <i>Drosophila yakuba</i> genome illustrating the boundaries of the three classes of the categorical variable (D)	63
Figure 4.5. Path diagram for 19 log-linear models fit to the <i>Drosophila yakuba</i> data	73
Figure 4.6. Path diagram for 19 log-linear models fit to the <i>Anopheles gambiae</i> data	79
Figure 4.7. Path diagram for 19 log-linear models fit to the <i>Apis mellifera</i> data	83
Figure 4.8. Comparison of patterns of synonymous codon usage for the three insects	87
Figure 5.1. Relative frequency of each base at third positions of leucine and valine codons vs. relative frequency of the base at all fourfold degenerate sites . . .	101
Figure 5.2. Relative frequency of each base at third positions of arginine and glycine codons vs. relative frequency of the base at all fourfold degenerate sites . . .	102
Figure 5.3. Relative frequency of each base at third positions of serine and threonine codons vs. relative frequency of the base at all fourfold degenerate sites . . .	103

Figure 5.4. Relative frequency of each base at third positions of alanine and proline codons vs. relative frequency of the base at all fourfold degenerate sites . .	104
Figure 5.5. Plots of all pairwise comparisons of the relative frequencies of each base at fourfold degenerate third codon positions	107
Figure 5.6. Relative frequency of each base at all first codon positions vs. relative frequency of the base at all fourfold degenerate third codon positions for 400 mitochondrial genes from 31 taxa	109
Figure 5.7. Relative frequency of each base at all second codon positions vs. relative frequency of the base at all fourfold degenerate third codon positions for 400 mitochondrial genes from 31 taxa	110
Figure 5.8. 95% Confidence Intervals for the estimated slope of regressions of first or second codon position composition and fourfold degenerate sites for each gene	113
Figure 5.9. Estimated slope for first and second codon position regressions vs. relative rate of replacement	116
Figure 5.10. Histograms of estimated slope and R ² values from the regressions of individual amino acid frequencies	120
Figure 5.11. A visual representation of the relative slope and significance level of each individual regression	121
Figure 5.12. Shifts in usage of twofold and fourfold degenerate codons for leucine	122

ABSTRACT

PATTERNS OF BASE COMPOSITION WITHIN AND BETWEEN ANIMAL MITOCHONDRIAL GENOMES

by

Nicole T. Perna

University of New Hampshire, September, 1996

Nucleotide composition of a DNA molecule is a product of base substitution. Variation in nucleotide composition indicates a change in the pattern of substitution at either the level of the underlying mutational spectrum or the constraints imposed by natural selection. This work explores patterns of nucleotide usage within and between animal mitochondrial genomes and the evolutionary mechanisms that have shaped these patterns. Fourfold degenerate sites are expected to reflect the underlying mutational spectrum. Three simple measures of compositional bias, taking into account the strand-specific nature of nucleotide distribution in mtDNA, reveal considerable variation among fourfold degenerate sites of metazoan mitochondrial genomes. Log-linear analysis of intramolecular compositional patterns of mammalian mtDNA demonstrates that fourfold degenerate sites from even a single strand of the genome are not homogeneous. Rather, base composition varies among codon families and around the circular genome. A companion analysis of two additional taxonomic groups, molluscs and insects, also reveals compositional variation among codon families and between strands. The observed intramolecular variation cannot be explained solely by a simple strand-specific mutational pressure, but requires either a contextual bias to the mutational process or translational level natural selection as well. First and second codon position base composition and amino acid frequencies regressed on fourfold degenerate site composition show how mutational biases at the DNA level translate to amino acid biases in mitochondrial proteins.

CHAPTER I

INTRODUCTION TO BASE COMPOSITION, UNEQUAL BASE SUBSTITUTION AND PHYLOGENETIC RECONSTRUCTION

DNA was originally dismissed as a candidate for the agent of heredity because of its apparently simple composition. The means by which such a simple molecule carries so much information, what this information is, and how it is organized, expressed and evolves has been a major focus of biological research for half a century. At the time of the observation that the transmission of DNA is the transmission of information, it became clear that the components are not randomly organized. From Chargaff's (1950) rules that played a critical role in the elucidation of DNA structure, to the current onslaught of genome projects, we have been cataloging and deciphering patterns of nucleotide composition.

The simple components of DNA are organized into four nucleotides, each composed of a deoxyribose covalently linked to a phosphate and one of the nitrogenous bases, guanine (*G*), adenine (*A*), thymine (*T*) or cytosine (*C*). The double helical structure of DNA in living cells was described by Watson and Crick in their well known 1953 paper *The Structure of DNA*. A single strand of DNA is a polymer of nucleotides linked by phosphodiester bonds. Two anti-parallel strands are held together by hydrogen bonding between *A* and *T* and between *G* and *C*. Thus, the first fundamental constraint on the base composition of a double-stranded DNA molecule is equimolar ratios of complementary nucleotides. Other than this structural limitation, it appears that the nucleotide composition of DNA sequences is entirely free to vary.

Of course, the DNA sequence of a living organism does not arise anew but has been

evolving since a single distant ancestor lived roughly a billion and a half years ago. Thus, a second restriction on the order and number of nucleotides in a DNA molecule is that it must have evolved. Genomes evolve as a multitude of mutational mechanisms introduce genetic variation. Sequence variants with no current appreciable phenotypic consequence will be maintained or lost through genetic drift. When phenotypic variation affects fitness, natural selection can affect the fate of the variant. As species emerge and diverge, they experience different patterns of mutation and selective pressures. These evolutionary forces lead to compositional variation among genomes as a whole and even among homologous genes. Understanding patterns of composition is central to using DNA to reconstruct evolutionary history of organisms, as well as to reconstructing the evolutionary history of an organism's DNA. Nowhere is this reciprocal relationship more apparent than in studies of mitochondrial DNA (mtDNA).

Mitochondrial DNA Evolves by Base Substitution

The mitochondrial genome has been a workhorse of animal evolutionary genetics for the past decade and a half. Uniparental inheritance and a rapid rate of evolution make this molecule well suited for revealing population structure and the systematics of recently diverged taxa. With the development of molecular phylogenetic analysis using mtDNA sequences, we have learned much about patterns and rates of evolution in mitochondrial genomes. In 1991, when the work reported herein was initiated, there were nine complete metazoan mitochondrial genome sequences published. By the time the last analysis in this thesis began, there were over thirty available in the nucleotide database maintained by NCBI (National Center for Biotechnology Information). Several more have been added since. The rapid acquisition of these genome sequences is a function of both ease and interest. Complete genome sequences are useful for studying mitochondrial molecular biology (Clayton 1982, 1992), as well as molecular evolution (Wolstenholme 1992).

Metazoan mitochondrial genomes are small and isolation is facilitated by the high copy number within cells. Despite their small size, these genomes contain a tremendous amount of information. Each circular genome of 12,000-20,000 base pairs encodes 12 or 13 messenger RNAs, 22 transfer RNAs and 2 ribosomal RNAs. The non-coding sequence between most genes is limited to a few base pairs. All metazoan mitochondrial genomes contain a major non-coding segment and the length variation among genomes is primarily the result of changes within this region. Those elements of the translational machinery encoded in mtDNA are the sole remnants of a larger set of housekeeping genes that must have been necessary in the free-living prokaryotic ancestor of the mitochondrion. Endosymbiotic invasion and subsequent streamlining of the genome (Margulis 1970) have produced the compact molecule we see today. Although even the basic processes of genome maintenance and expression require importation of additional gene products encoded in the nucleus, the protein products encoded in mtDNA are essential. All 13 protein-coding genes in animal mtDNA encode important components of oxidative phosphorylation machinery including subunits of the ATP synthetase, NADH dehydrogenase, the *b-c₁* and cytochrome oxidase complexes (table 1.1). Natural selection preserving the function of these critical proteins prevents most insertion or deletion of base pairs within these genes from becoming fixed in the genomes of species. Variation among homologous genes from the handful of complete genomes, or the ten thousands of individual gene sequences now available, is primarily the result of single base substitutions.

Table 1.1. Thirteen mitochondrially encoded proteins and the abbreviations used to represent them throughout this dissertation. Entries marked * are used only in figures where the larger abbreviation will not fit and the shortened version can not be misconstrued.

Protein	Abbreviation (s)
NADH dehydrogenase	
subunit 1	ND1
subunit 2	ND2
subunit 3	ND3, N3*
subunit 4	ND4
subunit 4L	ND4L, L*
subunit 5	ND5
subunit 6	ND6, N6*
cytochrome c oxidase	
subunit I	COI
subunit II	COII
subunit III	COIII
adenosine triphosphatase	
subunit 6	ATPase6, A6*
subunit 8	ATPase8, 8*
cytochrome b	Cyt b

Phylogeny Reconstruction, Models of Base Substitution, and Base Composition

Molecular phylogenetics is founded on the principle that evolutionary relationships can be inferred from the patterns of variation among homologous DNA or protein sequences. The idea is conceptually simple: sequences that are more alike are more closely related than sequences that are very different from each other. There are many approaches to reconstruction of molecular evolutionary relationships, some of which differ in underlying philosophy and some of which differ only by relatively minor changes in complex algorithms. One thing common to all approaches is that their actual application to real data necessitates making assumptions about how molecules change over time. Patterns of base

composition within and among taxa are useful for testing these assumptions about patterns of base substitution and for understanding the evolutionary pressures that shaped these patterns.

Focusing for the moment on distance-based phylogeny reconstruction, it is easy to see the importance of base composition. The object is to assign each pair of sequences in the analysis a single value of genetic distance based on the amount of divergence between them. Various strategies are then used to construct a tree illustrating relatedness based on the pairwise distance matrix. The simplest possible way to assign a distance value is to count the raw number of differences between the two sequences. When the total number of base substitutions among sequences since their divergence from a common ancestor is small, the probability of a change at any given site is also small. It will be unlikely that a single site has experienced more than one substitution. The total number of differences will increase linearly with divergence time assuming that substitutions accumulate in a pseudo-clock-like manner. The existence and nature of such a molecular clock has been the subject of much research, much debate and much confusion. Even setting this formidable topic aside, complications quickly arise as the total number of base substitutions increases over evolutionary time and some sites experience multiple substitutions. Counting the number of differences between sequences will underestimate the total amount of divergence because multiple substitutions are invisible: A new substitution erases the record of a previous substitution at that site. Sequence differences no longer show a linear relationship with time, but rather approach a saturation point after which no amount of base substitution can increase the number of pairwise differences. The saturation point is dependent on the rates and patterns of evolution (Irwin et al. 1991). In order to correct estimates of genetic distance to reflect the true amount of divergence, it is necessary to assume a stochastic model of base substitution.

These models consist of a four by four matrix describing the rates of substitution from a

base to each of the other three bases. The one-parameter model (Jukes and Cantor, 1969), with an equal rate for all types of substitutions, predicts equal frequencies of the four nucleotides at equilibrium. From the first few mitochondrial sequences on, it has been apparent that most mtDNA does not have an unbiased base composition. All types of substitutions do not occur at equal rates. Early comparisons of recently diverged mtDNA sequences (Brown et al. 1982) showed a ten-fold excess of transition (purine \leftrightarrow purine or pyrimidine \leftrightarrow pyrimidine: $A \leftrightarrow G$ or $T \leftrightarrow C$) substitutions over transversions (purine \leftrightarrow pyrimidine). The two-parameter base substitution model (Kimura, 1980) allows rate differences between transitions and transversions. However, this modification does not change the expected equilibrium frequency of the four bases, and thus is still not a realistic model of base substitution for mtDNA. Notice that the methodology is based on observing rate differences among substitutions in a group of sequences then incorporating these differences into a model of substitution to analyze patterns of differences in other sequences. This approach has led to significant advances in phylogenetic reconstruction methods, but it is important to understand that the rate differences detected may themselves be dependent on a particular phylogenetic reconstruction and all its associated biases. Two further complications of phylogenetic analysis related to base composition will be addressed before revisiting the idea of using models to infer rates to construct models.

We have already pointed out that base composition can vary among even homologous DNA sequences, and a significant portion of this thesis is dedicated to describing compositional variation among mitochondrial genomes. Since base composition is the result of the pattern of base substitution, we can infer that substitution patterns vary among those lineages that differ in composition. All distance-based phylogenetic reconstruction methods assume that base substitution is a stationary process. In other words, the observed base composition reflects the nucleotide frequencies at equilibrium; The same substitution matrix is acting along all lineages and has been since the divergence from the

common ancestor of all sequences under analysis. A sensitive method for revealing deviations from a stationary process shows that surprisingly small compositional variations among recently diverged taxa can lead to violation of this assumption (Rzhetsky and Nei, 1995).

Other reconstruction methods are also sensitive to compositional variation. Phylogenetic analyses based on parsimony as an optimality criterion are as popular as distance based phylogenies. In a parsimony analysis, the goal is to choose a branching order that minimizes the total number of changes required to go from the common ancestor to all extant sequences. When composition varies among sequences, parsimony has a tendency to cluster sequences with similar base compositions. This becomes problematic when distantly related sequences have more similar base composition than sequences that share a more recent common ancestor. This situation can arise because of a change in the substitution matrix in a terminal branch of the phylogenetic tree or because similar substitution matrices have evolved independently in different lineages.

The pattern of base substitution varies not only among sequences, but also within a sequence. Underlying the process of substitution is a mutational spectrum created by misincorporation of nucleotides by DNA polymerases (Kunkel 1985) and spontaneous chemical degradation (Lindahl 1993). Base mismatches created by these factors may then be resolved by DNA repair mechanisms or lead to mutations. Finally, this collection of mutations is filtered by selection for function at either the level of the DNA or the product it encodes. Although we generally think of mutation as a process that is constant within a genome, there are precedents for mutational variation within a DNA molecule. One example is a dependence of substitution patterns for a given site on the adjacent nucleotides, known as the neighboring base effect (Bulmer 1990). Larger segments of a continuous DNA sequence can experience different patterns of mutation if they are replicated by different DNA polymerases, or in the presence of different concentrations of

the free nucleotides used to assemble the new strand of DNA (Wolfe et al. 1989). Certainly, there are numerous reasons to expect natural selection to impose different constraints on substitution patterns at different sites. The first, second and third codon positions of genes experience different selective pressures. The first and second codon position are constrained by selection for amino acids. Nucleotide usage at these positions must correspond to codons which lead to an acceptable protein. Just what constitutes an acceptable amino acid will vary between codon sites and is likely to be extremely limited in some regions of a protein, such as active sites, or may be flexible enough to encompass any amino acid with particular physical or chemical properties (hydrophobicity, charge, size) in some regions. Nucleotide sequences of tRNA and rRNA genes are constrained to encode products that fold into appropriate secondary and tertiary structures and function in translation. Observations on the heterogeneity among sites of the major non-coding region in mtDNA has had the greatest impact on how substitutional variation within a sequence is accommodated in phylogenetic reconstruction.

The major non-coding region of animal mitochondrial genomes is also known as the control region or displacement loop (D-loop) (Wolstenholme 1992, Clayton 1992). The role of the control region DNA sequence in the initiation of replication and transcription of the genome ensures that natural selection can affect the rates and patterns of evolution in this region. Some sites in the control region, however, are not involved in genome maintenance and regulation. These sites are undoubtedly among the most rapidly evolving sites in the genome. For this reason, the control region was an obvious candidate for resolving recent phylogenetic relationships, like reconstruction of human evolutionary history.

Nearly everyone has heard of mitochondrial Eve, the maternal ancestor of all modern humans. Eve was placed in Africa roughly 238,000 years ago on the basis of a phylogenetic reconstruction using control region sequences (Vigilant et al. 1991).

Although Eve's story provoked a myriad of criticisms for an equally diverse number of reasons, the most scientifically interesting of these criticisms are related to patterns of base substitution in the control region of hominoids. In order to reconstruct the timing of divergence from a common ancestor, sequence divergence within humans must be compared with sequence divergence from our closest living relative, the chimpanzee. This divergence is old enough (roughly 5 million years) to allow multiple substitutions to occur at some sites, and so the reconstruction requires a correction of sequence difference to estimate sequence divergence. Use of an inappropriate model of substitution to make this correction could affect both the inferred geographic origin and timing of the divergence. The observation that hominoid control region evolution was rapid at some sites and extremely slow at others (Kocher and Wilson 1991) meant that the most appropriate correction for multiple substitution would accommodate this rate heterogeneity. One approach to accommodating this heterogeneity among sites is to allow the rate of evolution to vary continuously according to a gamma distribution (e.g. Tamura and Nei, 1993). Allowing a gamma distributed rate provides a theoretical distribution of substitutions per site consistent with the negative binomial distribution observed by Kocher and Wilson (1991). This modification can now be added to any of a number of substitution models in the currently popular analysis program, MEGA. But what of Eve? Generations of substitution models later, she is still African, but considerably more modern.

The following discussion of substitution patterns and the sensitivity of divergence estimates to the assumed equilibrium base composition (pages 8-11) was published under the title "Unequal base frequencies and the estimation of substitution rates" (Perna and Kocher 1995a).

Tamura and Nei (1993) presented a new method for estimating the number of nucleotide substitutions between two sequences and demonstrated its applicability by reanalyzing the human D-loop data. Their model of substitution allows different rates of purine transition,

pyrimidine transition and transversion. It also includes the gamma rate variation modification to accommodate heterogeneity among sites. In order to maintain an equilibrium composition consistent with the base composition of human control regions, each substitution rate in the four by four matrix is weighted by the frequency of the mutant base. The motivation for this weighting arises from an analysis of the patterns and relative rates of base substitutions inferred by parsimony from a distance-based tree.

Table 1.2 shows the substitutions inferred by Tamura and Nei from 95 human control region sequences. Surprisingly, this matrix suggests that the base composition of the control region is changing over time. The number of *G*'s lost by substitution to another nucleotide is 41.5, while the number *G*'s created by mutation is 68.5. This suggests a net gain of 27 *G*'s over this phylogeny. A change in base composition seems unlikely for two reasons. First, nucleotide composition is conserved among hominoid mtDNA sequences over a time period considerably greater than that represented by this data (Kondo et al. 1993). Second, the pattern of mtDNA nucleotide composition in primates, and indeed throughout the animal kingdom, is characterized by a low frequency of *G* on this strand. If the substitution matrix inferred by parsimony is correct, then the base composition of these sequences is not at equilibrium. Evolving according to the inferred matrix, the composition of these sequences would eventually come to equilibrium at 0.26, 0.25, 0.32 and 0.18 for *A*, *T*, *C*, and *G* respectively. The observed composition is 0.32, 0.23, 0.31, and 0.13. It seems unlikely that the composition of human mitochondrial genomes are becoming more even than that of their ancestors, by a mechanism that reverses the directional pressure against *G* found in all known animal mtDNA.

Table 1.2. Substitution matrix for the human mitochondrial control region estimated by Tamura and Nei (1993).

MUTANT NUCLEOTIDE	ORIGINAL NUCLEOTIDE				TOTAL
	T	C	A	G	
T		115	2	2	119
C	112		5	2	119
A	1	5		37.5	43.5
G	1	3	64.5		68.5
TOTAL	114	123	71.5	41.5	350

The inferred compositional shift is probably an artifact created by assumptions used in reconstructing ancestral states. Figure 1.1a shows the pattern of substitution inferred for three terminal character states, using the method which produced the Tamura and Nei matrix. If we assume that the composition of the variable sites is equal to the control region as a whole, then the probability of transition $G \rightarrow A$ must be 2.4 times higher than the probability of transition $A \rightarrow G$, in order to maintain this equilibrium composition (figure 1.1b). Under these conditions, the alternative character state reconstruction shown in figure 1.1c is likely to represent the true pattern of substitution at some sites. By failing to account for unequal frequencies of nucleotides in the sequence, and hence unequal rates of substitution, simplistic applications of parsimony may incorrectly reconstruct ancestral character states, causing an underestimation of the total number of substitutions.

The probabilities of forward and backward transition substitution (figure 1.1b) may be even more unequal than Tamura and Nei estimated. In applying their model, they used the average frequency of each nucleotide at all sites in the control region. Base composition varies, however, depending on the degree of selective constraint on a site. Fourfold degenerate third codon positions are among the least constrained positions of the

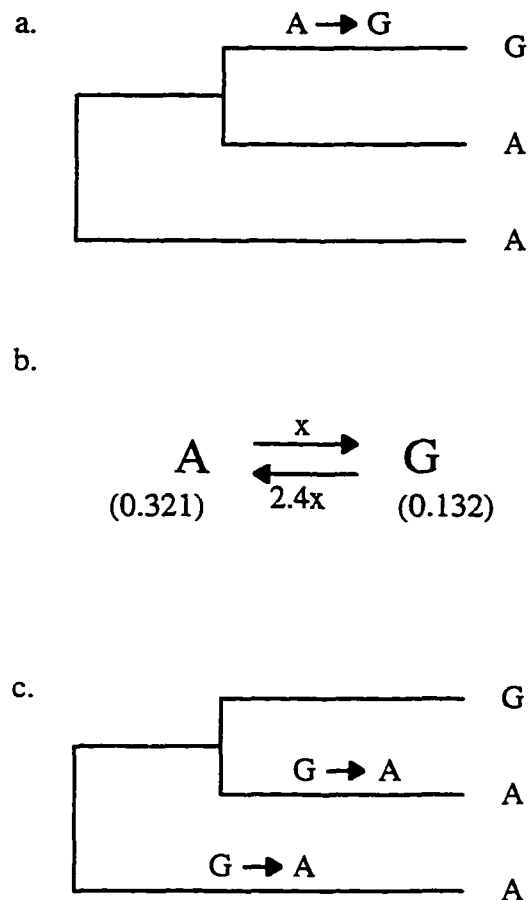
mitochondrial genome and base frequencies at these sites are much more unequal than those found in the control region taken as a whole. Selective forces acting on the control region (Kocher and Wilson 1991) may constrain base frequencies at some sites. The heterogeneity of selective constraints will be reflected as differences in equilibrium nucleotide composition among rapidly and slowly evolving sites.

We suggest that using the average composition of the control region as an estimate of equilibrium nucleotide frequency causes an underestimation of the substitution inequalities actually present. If the composition of the most variable control region sites is similar to that of fourfold degenerate sites (where the frequency of A=0.402 and G=0.054) the actual base-specific rate of G→A substitution may be 7.4-fold greater than the rate of A→G substitution. Since the parsimony method used by Tamura and Nei may have greatly underestimated the number of purine transitions, their conclusion that the pyrimidine transition rate is higher should be reexamined.

While it is not clear that either the average composition of the control region or fourfold degenerate sites truly represent the equilibrium frequencies of nucleotides at variable sites in the control region, it is interesting to compare the results arising from different assumptions. We have reanalyzed the complete human and chimp control region sequences using Tamura and Nei's model of substitution. We compare the results obtained using the overall composition of the control region, with those obtained using base frequencies at fourfold degenerate sites. The frequencies of A, T, C, and G at the fourfold sites are 0.402, 0.132, 0.411, and 0.054 respectively. The average estimate of genetic distance, d , among humans is similar using both the control region (0.024 ± 0.006 , Tamura and Nei 1993) and fourfold site (0.030 ± 0.009) compositions. However, the choice of equilibrium base composition has a much greater effect on the average d for the more divergent chimpanzee sequences used to estimate the modal substitution rate. If we assume a divergence time of 5 MY between humans and chimps, we calculate a modal divergence

rate of 2.13×10^{-7} using the fourfold site composition. This is much higher than the rate calculated from the average D-loop composition (7.5×10^{-8}) and would force a revision of the estimated age of the common human ancestor from 160,000 years to just 71,000 years.

Figure 1.1. Inferring directed matrices of substitution from observed character states using parsimony criteria. The simplest parsimony reconstruction (a) assumes an equal probability of A→G and G→A substitutions. If the frequency of A and G are unequal, the probability of substitution will be strongly asymmetric (b). In this case, the alternative reconstruction (c) is likely at some sites.



Studies employing the principle of parsimony to infer patterns of substitution need to address the effect of systematic reconstruction biases among even closely related sequences

when nucleotide frequencies are unequal. We have observed this same non-stationary phenomenon in several other published studies (Palumbi and Kessing 1991; Tamura 1992; Knight and Mindell 1993). Another group (Collins et al. 1995) has shown that this tendency of parsimony to incorrectly infer a compositional shift becomes more exaggerated with increasing sequence divergence. We expect that the stationary model of substitution developed by Tamura and Nei, when properly applied, will be one of the best methods to estimate divergence for sequences in which the four nucleotide frequencies are unequal. Accounting for unequal nucleotide composition in substitution models is not a trivial matter pertaining to only a few data sets. Most mitochondrial, prokaryotic and many nuclear genomes exhibit some compositional inequalities. Analyses of ribosomal DNA, in which selective constraints and base composition vary among sites, are likely to present problems very similar to those encountered in the analysis of the mitochondrial control region.

Beyond phylogenetic reconstruction

The addition of a gamma parameter to substitution models greatly improves the accuracy of divergence estimates by incorporating the observation that different sites evolve at different rates, but assumes that the pattern of substitution is constant for all sites. If base composition is at equilibrium for all sites, yet varies among subsets of the molecule, then clearly, a single pattern of evolution can not be realistic. In theory, a model of base substitution could include a separate transition probability matrix for each site in a group of aligned sequences. This would allow adjustment not only for rate heterogeneity, but also for differences in patterns of substitution among sites. However, such a model would not be useful because of the large error associated with estimating so many parameters from the data. It is desirable for statistical models of base substitution to reflect the true process, but clearly it is necessary to draw a line somewhere and call a model realistic enough. The placement of this line is largely determined by the goal of the phylogenetic reconstruction.

Careful selection of a molecule evolving at the correct rate for the taxa being analyzed can reduce the complexity of the model required for analysis of the data.

The remainder of this dissertation is dedicated to improving a conceptual model of base substitution in mtDNA without regard for whether or not these observations can, should or must be incorporated into the statistical models used to reconstruct phylogeny. Chapter II describes overall patterns of compositional variation among animal mitochondrial genomes, taking into account the strand-specific nature of nucleotide distribution in mtDNA. Chapter III is a more in depth analysis of intramolecular compositional patterns in mammalian mtDNA, and Chapter IV is a companion analysis of two additional taxonomic groups: molluscs and insects. Intramolecular variation described in these chapters reveals complex evolutionary pressures acting in even the simplest subset of mtDNA sites. The final chapter investigates how mutational biases at the DNA level translate to amino acid biases in mitochondrial proteins.

CHAPTER II

STRAND-SPECIFIC PATTERNS OF NUCLEOTIDE COMPOSITION AT FOURFOLD DEGENERATE SITES OF ANIMAL MITOCHONDRIAL GENOMES

Three statistics (*%GC*, *GC-skew* and *AT-skew*) can be used to describe the overall patterns of nucleotide composition in DNA sequences. Fourfold degenerate third codon positions from 16 animal mitochondrial genomes were analyzed. The overall composition, as measured by *%GC* varies from 3.6 *%GC* in the honeybee to 47.2 *%GC* in human mtDNA. Compositional differences between strands of the mitochondrial genome were quantified using the two skew statistics presented in this paper. Strand-specific distribution of bases varies among animal taxa independently of overall *%GC*. This chapter has been previously published (Perna and Kocher 1995b).

Introduction

The nucleotide composition of mitochondrial genomes varies among animal taxa. For example, the complete mitochondrial DNA sequence of the honeybee is only 15.1% *GC* base pairs (Crozier and Crozier 1993) while that of the human (Anderson 1981) is 44.4% *GC*. Compositional differences also exist between the two strands of the mitochondrial genome and were originally recognized as differences in buoyant density in CsCl gradients (Brown 1981). The biochemical and evolutionary origins of these compositional features and the relationship between the strand-specific distribution and the overall *%GC* of the genome, are presently unknown. Clearly, nucleotide usage results from the process of substitution, but many of the factors which affect the pattern and rate of substitution in

mtDNA are not well characterized.

Underlying the process of substitution is a mutational spectrum created by misincorporation of nucleotides by DNA polymerases (Kunkel 1985) and spontaneous chemical degradation (Lindahl 1993). Base mismatches created by these factors may then be resolved by repair mechanisms or lead to mutations. Finally, this collection of mutations is filtered by selection for function at either the level of the DNA or the product it encodes. Thus, patterns of composition within genomes and compositional differences among homologous sequences could result from both variation in the selective constraints and changes in the mutational spectrum during evolutionary divergence.

In this paper we use three measures to describe nucleotide patterns at fourfold degenerate third codon positions, because of all sites in the mitochondrial genome, these are most likely to reflect the underlying mutational matrix. In studies of nuclear genomes, non-coding or pseudogene sequences are often used to study the mutational matrix (for example Bulmer 1985). Unfortunately, the major non-coding region of mtDNA has important, if poorly understood, functions which exert strong selective constraint (Kocher and Wilson 1991). Likewise, patterns of selection on tRNA and rRNA genes arising from secondary structure and interactions with other molecules are complex, and it is difficult to define a homogeneous subset of sites from these genes (Xiong and Kocher 1993). First and second codon positions are subject to selection for amino acid sequence in the resulting protein, and thus are not good estimators of compositional patterns generated by the mutational spectrum. Some studies of mtDNA composition (e.g. Asakawa et al. 1991) have included all third codon positions. Mitochondrial proteins have highly biased amino acid composition and unequal numbers of two-fold degenerate codons, which may affect estimates of equilibrium composition calculated from all third positions.

Although fourfold degenerate sites are free of the selective constraints of amino acid specification, composition at these sites may still be affected by selection for translational

efficiency. Synonymous codon usage in bacteria and yeast is strongly correlated with overall composition of the genome, yet highly expressed genes often use a higher proportion of certain codons to promote efficient translation (Shields and Sharp 1987). Selection to match codon-usage with iso-accepting tRNA abundance is unlikely to be important in the mitochondrial system where there is usually only one tRNA for each fourfold degenerate codon family. However, there might be selection among synonymous codons for different binding affinities to the tRNA anticodon (Bulmer 1991). Asakawa et al. (1991) argue that this is also unlikely to be a factor in mitochondrial composition because mitochondrial genome rearrangements have periodically caused some genes to switch strands. These genes evolve base compositions consistent with their new location rather than their original strand.

If the composition of fourfold degenerate sites is primarily the result of the mutational matrix, then it is clear that the directional mutation pressure is strand-specific, unlike the directional mutation pressure modeled by Sueoka (1988) to address variation in %GC among eukaryotic nuclear and bacterial sequences. We hope that an accurate and quantitative description of compositional patterns at fourfold degenerate sites, and an exploration of the evolutionary history of compositional variation in mtDNA, will provide insight into the nature of the mutational pressures acting on this molecule.

Methods

Sequences

Complete mitochondrial genome sequences are available for thirteen taxa included in this analysis: *Apis mellifera* (Crozier and Crozier, 1993), *Ascaris suum* (Okimoto et al. 1992), *Bos taurus* (Anderson et al. 1982), *Caenorhabditis elegans* (Okimoto et al. 1992), *Crossostoma lacustre* (Tzeng et al. 1992), *Cyprinus carpio* (Chang et al. 1994), *Drosophila yakuba* (Clary and Wolstenholme 1985), *Gallus gallus* (Desjardins and Morais 1990),

Homo sapiens (Anderson et al. 1981), *Mus musculus* (Bibb et al. 1981), *Paracentrotus lividus* (Cantatore et al. 1989), *Petromyzon marinus* (Lee and Kocher, 1995), and *Strongylocentrotus purpuratus* (Jacobs et al. 1988). For these taxa, we used fourfold degenerate codon positions from all the mitochondrial protein coding genes which are encoded on one strand of the genome, known as the heavy strand in vertebrates (Brown 1981). The motivation for using only genes encoded on the same strand is to minimize the potential confounding compositional effects of different evolutionary pressures experienced by sequences transcribed at different rates (Attardi et al. 1982) and replicated asymmetrically (Clayton 1992). In nematodes, the replication mechanism is unknown and cannot be inferred by phylogenetic comparison, but all twelve protein coding genes are on the same strand and are included in this analysis. At the time of this analysis, complete mitochondrial genome sequence was not available for any mollusc, so we have filled this phylogenetic gap with sequences from twelve *Mytilus edulis* genes, all of which are encoded on the same strand (Hoffmann et al. 1992). We have also included data from partial genome sequences for two additional echinoderm taxa, *Arbacia lixula* (De Giorgi et al. 1991a) and *Asterina pectinifera* (Himeno et al. 1987) because the replication mechanism differs between urchins and sea stars and a strand-specific mutation pattern has been observed in *Arbacia* (DiGiorgi et al. 1991b). Partial ND5 (189aa) and COIII (109aa) sequences were used for *A. lixula*. The *A. pectinifera* data come from partial COIII (69aa) and ND5 (512aa) and complete ND3 and ND4 sequences.

Codon frequencies

The composition of fourfold degenerate third codon positions of the 16 species was calculated by generating codon frequency tables with the GCG CODONF program (Devereux et al. 1984) and summing the frequency of each base at the third positions across the 8 fourfold degenerate codon families (glycine-*GGN*, leucine-*CTN*, valine-*GTV*,

arginine-*CGN*, threonine-*ACN*, alanine-*GCN*, serine-*TCN*, and proline-*CCN*) common to all variations of the animal mitochondrial genetic code. The *CTN* leucine codons were included even though they are actually 6-fold degenerate. The additional two-fold degeneracy results from synonymous changes at the first codon position. A large proportion of the total fourfold degenerate codons found in mitochondrial proteins are *CTN* leucines. We were concerned that first position substitutions in the two-fold degenerate *TTR* leucine codons might inflate the number of codons from the fourfold family which end in A and G. A preliminary analysis performed excluding the leucine codons suggested that this would not be a problem.

Statistics

We calculated three measures of compositional distribution from the fourfold degenerate third codon position nucleotide frequency data. Complementary pairing of bases permits all three to be calculated from the frequencies of nucleotides on a single strand.

The overall composition of the double-stranded molecule is measured by the proportion of *G+C* out of the total. This is a commonly used measure most frequently and simply described as %*GC*. The other two measures describe the compositional difference between the two strands:

$$GC-SKEW = (G-C)/(G+C) \quad (1)$$

$$AT-SKEW = (A-T)/(A+T) \quad (2)$$

where *G*, *A*, *T* and *C* are the frequencies of each nucleotide from the sense strand.

These two equations differ from other measures of strand-specific composition in several ways. The bias statistic used by Irwin et al. (1991) measures the deviation of the four bases from equal frequency, confounding %*GC* with strand-specific compositional patterns. Thomas and Wilson (personal communication) have recommended the statistic:

$$Skew = 2|G| - C| + 2|A| - T| \quad (3)$$

where *GI*, *CI*, *AI* and *TI* are proportions of each nucleotide on a single strand of the DNA helix. The most problematic feature of this statistic is that it confounds two aspects of skew, that arising from *AT* pairs with that involving *GC* base pairs. Numerically equivalent values of this statistic can arise from very different patterns of skew. We feel that it is important to consider the contribution of each type of nucleotide pair to the overall skew. Also, it is usually most convenient to express such statistics within a range from 0 to 1. We have also chosen to standardize our skew measures by the composition of the double-stranded molecule in order to consider different, possibly independent, patterns of composition. Finally, we have eliminated the absolute value bars to distinguish the direction of each type of skew. These measures are similar to those used by Saccone et al. (1993), differing only by removal of absolute value bars in the numerator. It is important to note that the sign of the skew statistic is meaningful only with reference to a particular strand since the same values with the opposite sign describe the composition of the other strand. However, the sign will provide an additional level of discrimination in comparative analyses where we can unambiguously identify one strand with reference to an asymmetric biochemical process, such as the direction of replication.

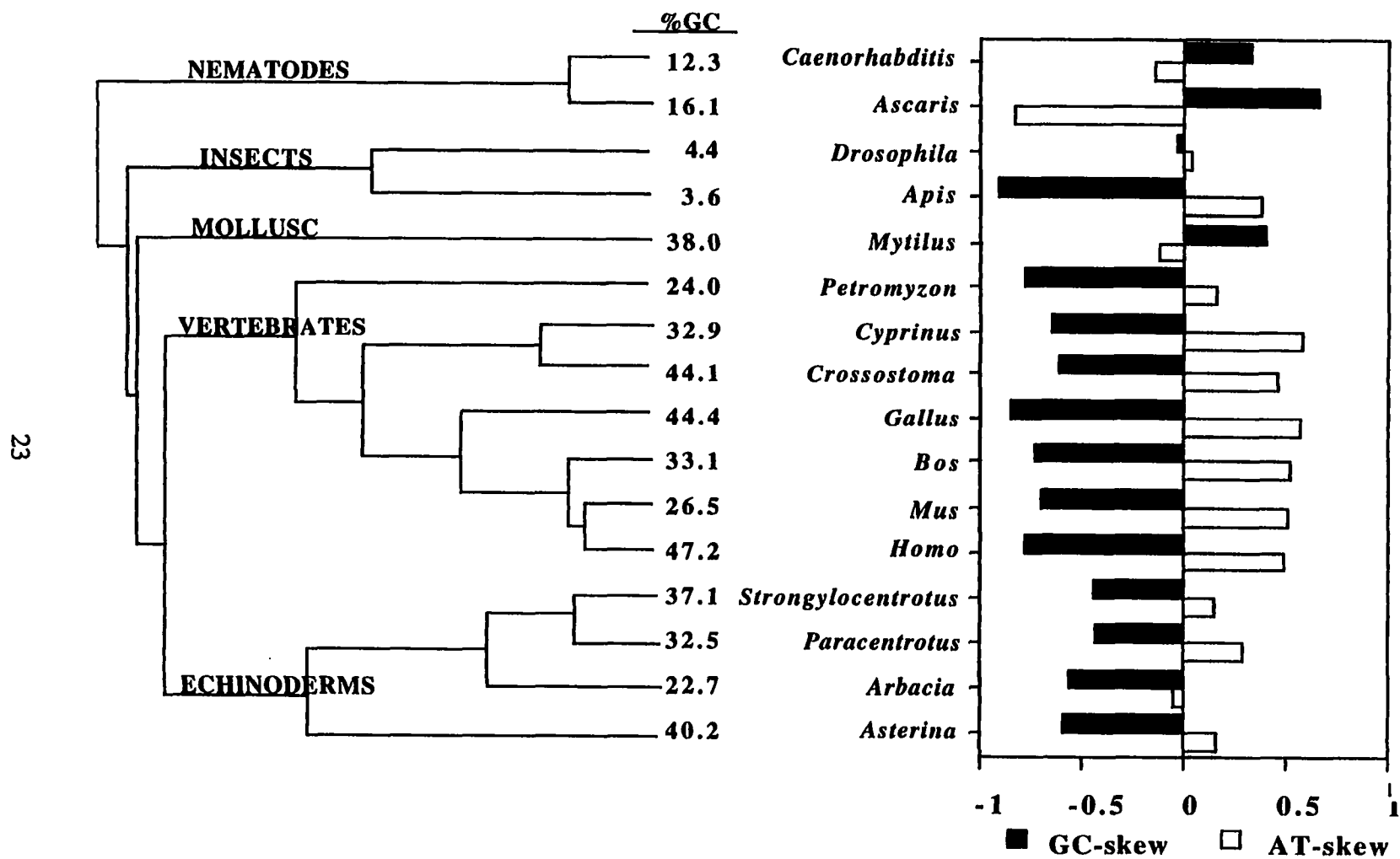
Results and Discussion

Base composition for the fourfold degenerate sites from 16 taxa is reported in Table 2.1. These nucleotide frequencies were used to calculate the %GC and values for both skew measures found in Figure 2.1. It is useful to consider these characteristics in a phylogenetic context. Examination of the phylogenetic distribution of compositional patterns allows the formation of hypotheses about which lineages have experienced changes in the substitution process. The phylogeny is derived from a combination of morphological and molecular evidence, and represents a consensus of current opinion on the relationship of the taxa used in this study (see Sidow and Thomas 1994).

Table 2.1. Nucleotide composition of fourfold degenerate third codon positions from 16 animal taxa.

<i>Genus species</i>	%G	%A	%T	%C
<i>Caenorhabditis elegans</i>	8.2	37.5	50.2	4.1
<i>Ascaris suum</i>	13.4	7.1	77.6	2.7
<i>Drosophila yakuba</i>	2.2	49.5	46.1	2.3
<i>Apis mellifera</i>	0.2	66.7	29.7	3.4
<i>Mytilus edulis</i>	26.5	27.4	34.6	11.4
<i>Petromyzon marinus</i>	2.7	44.2	31.8	21.3
<i>Cyprinus carpio</i>	5.7	53.0	14.2	27.1
<i>Crossostoma lacustre</i>	8.5	40.7	15.2	35.5
<i>Gallus gallus</i>	3.3	43.6	12.0	41.1
<i>Bos taurus</i>	4.5	51.0	16.0	28.6
<i>Mus musculus</i>	4.0	55.6	17.9	22.6
<i>Homo sapiens</i>	5.1	39.4	13.4	42.1
<i>Strongylocentrotus purpuratus</i>	10.3	36.3	26.6	26.8
<i>Paracentrotus lividus</i>	9.3	43.4	24.1	23.2
<i>Arbacia lixula</i>	4.9	36.6	40.7	17.7
<i>Asterina pectinifera</i>	8.2	34.7	25.1	31.9

Figure 2.1. Phylogenetic distribution of %GC ,GC-skew and AT-skew.



Nematodes:

Although the two nematodes, *C. elegans* and *A. suum* have a very similar overall composition (12.3 and 16.1 %GC) there are dramatic differences in the way the ~85% A's and T's are distributed on the two strands of each genome. In *C. elegans*, the sense strand has 37.5% A and 50.2% T, whereas in *Ascaris*, the same strand has only 7.1% A and 77.6% T. *Ascaris* has a much stronger AT-skew than *C. elegans*. The GC-skew of these two taxa is similar. Okimoto et al. (1992) suggested that *A. suum* and *C. elegans* may have shared a common ancestor as recently as 80 MYA. If this is true, the striking difference in AT-skew between these two taxa has arisen in a relatively short time. However, this divergence time is based on an estimated number of substitutions and an assumed rate of divergence derived from mammalian mtDNA studies. Because there have obviously been changes in the process of substitution along these lineages, as evidenced by the difference in AT-skew, this correction of sequence distance to divergence may be inadequate.

Insects:

The *D. yakuba* (4.4 % GC) and *A. mellifera* (3.6 % GC) fourfold degenerate sites are even more AT rich than the nematode mtDNA. *Drosophila* mtDNA is the least skewed of all genomes considered here. Both GC-skew and AT-skew are small in this sequence and not significantly different from zero. The honeybee genome is more skewed than the fruitfly genome for both types of base pairs. Only one fourfold degenerate third codon position containing G was found in the *Apis* sample, compared to fourteen sites containing C. These two insect taxa diverged from each other approximately 280MYA (Crozier and Crozier 1993). Although both nematodes and insects have very AT rich genomes, the skew patterns within and between these taxonomic groups are quite different.

Molluscs:

This group is represented solely by the *Mytilus* sequence, which is 37.8% GC. Little AT-skew is observed in this genome; however, the GC-skew is of similar magnitude and direction to that found in nematodes. Thus far, this pattern is seen only among invertebrates, but is not a universally conserved feature.

Vertebrates:

The %GC varies among vertebrates from 24.0% in lamprey, *Petromyzon marinus*, to 47.2% in human, *Homo sapiens*. Yet, the negative GC-skew and positive AT-skew pattern is conserved in all these taxa. There is relatively little variation in the magnitude of either type of skew among vertebrate mitochondrial genomes, with the exception of a reduced AT-skew in lamprey mtDNA. GC-skew varies from approximately -0.65 in the two teleost fish to approximately -0.85 in the chicken. AT-skews varies from 0.47 to 0.58 among vertebrates other than the lamprey (AT-skew=0.16).

Echinoderms

Echinoderm genomes are less skewed overall than vertebrate mtDNAs. However, there is a substantial negative GC-skew in all four echinoderm taxa. *Arbacia lixula* mtDNA has less AT-skew than other echinoderm sequences and the direction of this skew differs from the common deuterostome pattern. This is primarily due to the increased %T in *Arbacia*. Observed asymmetries in the *Arbacia* substitution matrix (De Giorgi et al. 1991b) involve A \leftrightarrow G transitions on the strand used in our analysis and do not explain the high %T.

Correlations among the statistics:

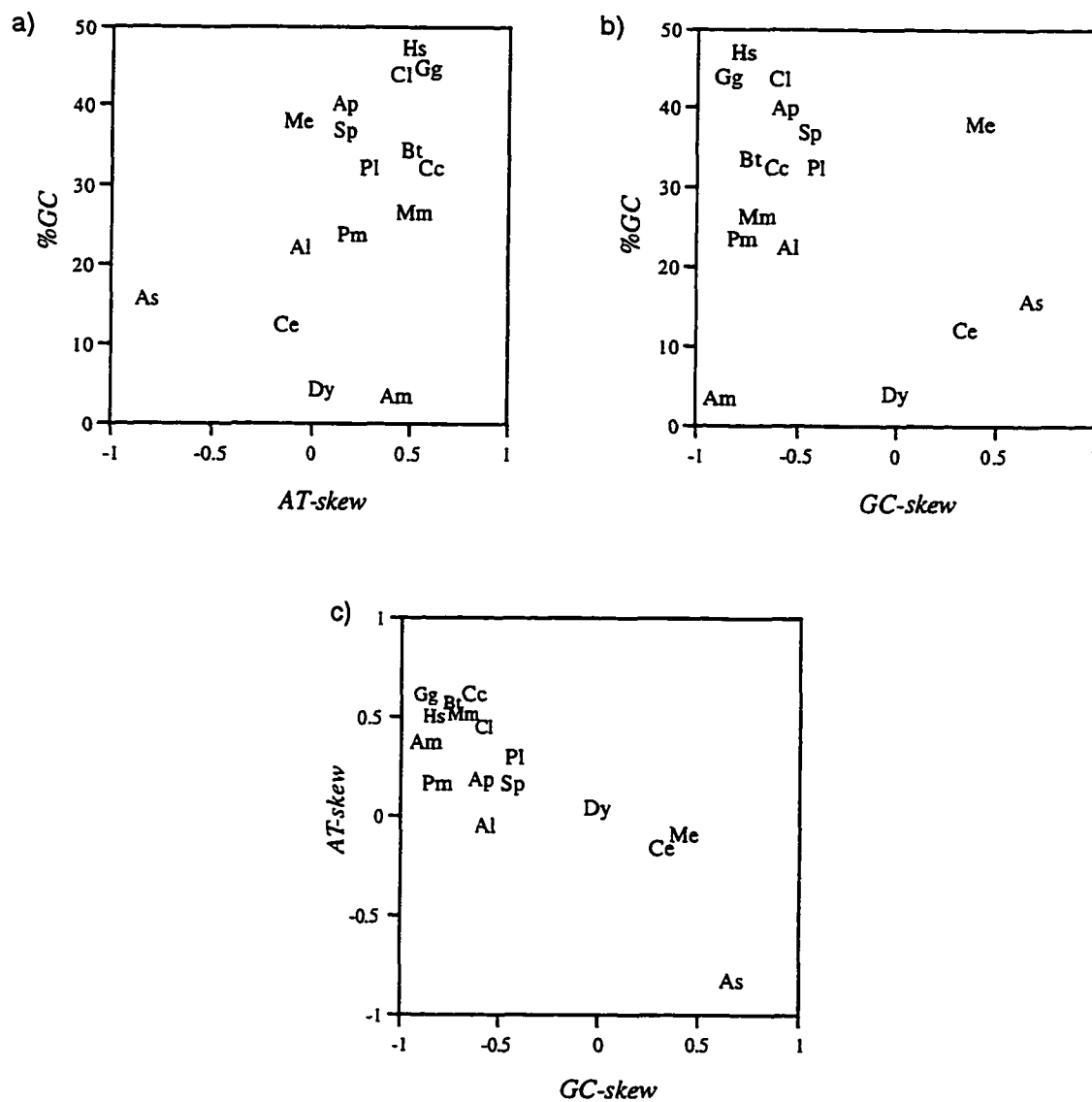
Although the three statistics are logically independent, they are related by a common substitution matrix. Correlations might be expected to arise, and may provide insight into

the mechanisms generating variation in overall %GC and skew. At the very least, scatterplots (Fig. 2) help to identify genomes with unique patterns of base composition. There is little correlation between %GC and either of the skew values (Figure 2.2a and 2.2b), which is consistent with the idea that separate mechanisms control the overall composition of the double stranded molecule and the strand-specific distribution of nucleotides (Thomas and Wilson, personal communication). We have already pointed out two instances where taxa with similar %GC have had dramatic differences in skew (nematodes and insects). Note also that the vertebrate genomes, while possessing a wide range of %GC, all have similar skews.

The relationship between *GC-skew* and *AT-skew* is shown in Figure 2.2c. No sequence considered here exhibited highly positive values for both *GC-skew* and *AT-skew*, or highly negative values for both skew measures. There is an interesting phylogenetic structure in the distribution of points in Figure 2.2c. Vertebrates cluster in the upper left quadrant, echinoderms cluster around 0.0 *AT-skew* and -0.5 *GC-skew* and other invertebrates are distributed in the region of positive *GC-skew*. The only exception is *Apis*, which clusters with the vertebrate taxa. *Ascaris* is the only sequence to exhibit a strong negative *AT-skew*.

Earlier we mentioned that all these genes are believed to be on the same strand relative to the direction of replication, but that for several of the taxa, direct experimental evidence is not currently available. If the strands are reversed in nematodes and *Mytilus*, a skew pattern would be common to all animal mtDNA's except *Arbacia* and *Drosophila*. The relationship between *AT-skew* and *GC-skew* shown in figure 2.2c would then disappear, leaving all three statistics largely uncorrelated. Molecular studies to confirm the mechanisms of replication of additional mitochondrial genomes would be useful.

Figure 2.2. Scatterplots are shown of a) *AT-skew* vs %GC, b) *GC-skew* vs %GC, and c) *GC-skew* vs *AT-skew*. Each point is represented by a two letter abbreviation corresponding to the first letter of the genus and the first letter of the species of the samples listed in table 2.1.



Variation within genomes:

An important question is whether the values of these compositional statistics vary within mitochondrial genomes. As we subdivide the mitochondrial genomes to examine patterns of composition within a molecule, the number of appropriate sites for analyses of this kind becomes quite small, especially for less frequent nucleotides. While some variation does exist among genes, a preliminary analysis indicates that intramolecular variation is small relative to the interspecific differences discussed in this paper. This is by no means intended to minimize the importance of intramolecular compositional patterns to understanding substitution mechanisms. Characterization of the overall patterns of composition is essential to the formation of testable hypotheses about the mechanistic origin of variation within genomes.

Predictions about the distribution of compositional variation both within and between genomes can be made from specific hypotheses about forces which shape the substitution process. For example, one hypothesis attributes strand-specific differences in the substitution matrix to differences in the damage spectra of single- and double-stranded DNA. This hypothesis was first put forward by Brown and Simpson (1982). *In vitro*, the rate of cytosine deamination is elevated approximately 200-fold in single-stranded DNA (Lindahl 1993). The asymmetric mechanism of replication (Clayton 1992) leaves some regions in a single-stranded state for as much as 30 minutes. The time spent single-stranded will vary in proportion to the distance from the replication origins (Thomas and Wilson, personal communication). This hypothesis has the potential to explain the common GC-skew pattern of vertebrates and echinoderms. Additional characterization of compositional patterns is necessary, however, to determine whether intramolecular compositional gradients consistent with this popular hypothesis exist.

Conclusions

It has long been recognized that the *GC* content of animal mitochondrial DNA varies widely among taxa, and that the composition of the two strands is not equal in higher vertebrates, particularly mammals. It has not been widely appreciated that the two kinds of base pairs (*GC* and *AT*) can have separate behaviors. The statistics presented here will be useful for quantifying this variation in mitochondrial and other genomes. A clearer description of the patterns of nucleotide composition may ultimately lead to a better understanding of the biochemical mechanisms creating the patterns.

The biochemical mechanisms of mutation, repair, replication, transcription and translation of mtDNA must all be taken into account as we search for the origin of compositional bias and skew in this molecule. Patterns of composition among individual genes and within fourfold degenerate codon families must be described in order to test a variety of hypotheses about composition and substitution, including the assumption that fourfold degenerate sites are not experiencing translational level selection. We are currently investigating the use of log-linear modelling to examine the relationship between a number of characteristics of the mitochondrial genome and base composition.

Finally, there is a need for additional biochemical studies. Most of the information we have about mitochondrial replication, polymerase specificity and DNA repair has come from studies of humans or mice. Studies of additional animal taxa are needed to provide comparative data, so that we may better understand the evolution of animal mtDNA.

CHAPTER III

LOG-LINEAR ANALYSIS OF SYNONYMOUS BASE COMPOSITION: MUTATIONAL BIASES AND TRANSLATIONAL LEVEL SELECTION IN MAMMALIAN MTDNA

Nucleotide usage at fourfold degenerate third codon positions varies among mammalian mitochondrial genomes. Deviations from equal frequency of synonymous codons and variations in codon usage among animal mitochondrial genomes are thought to result from directional mutation pressures. Mutational biases in mitochondrial DNA (mtDNA) are strand-specific. This log-linear analysis demonstrates that nucleotide composition varies around the molecule within a single strand of these mitochondrial genomes. Hypotheses which suggest that mutational pressures are related to the asymmetric mechanism of replication could account for both inter- and intra-strand compositional patterns. A codon family specific pattern of nucleotide usage detected in these genomes suggests that translational level selection may play an important role in synonymous codon usage.

Introduction

Synonymous codons are not used equally in animal mtDNA. Since the observation of a preponderance of codons ending with C and A in the human mitochondrial genome (Anderson et al. 1981), many researchers have reported base composition biases in third codon positions of mitochondrial sequences. While patterns of codon usage in many other systems, including the *E. coli* circular chromosome (Ikemura 1985), *B. subtilis* circular chromosome (Shields and Sharp 1987), yeast nuclear genomes (Sharp and Lloyd 1993; Lloyd and Sharp 1992), and *C. elegans* nuclear genome (Stenico et al. 1994), have been

attributed to both directional mutation pressures and translational level selection, the patterns observed in mitochondrial DNA have most often been explained by mutational pressures alone.

Evidence for strong mutational biases in mtDNA lies principally in the correlations between synonymous and nonsynonymous sites in the genome for both *GC* content (Jermiin et al. 1994) and strand-specific composition patterns (Asakawa et al. 1991). Several hypotheses about the origin of mutational biases are suggested by the asymmetric replication mechanism. Replication of one strand in mammalian mitochondrial DNA is initiated at an origin in the control region, the only major non-coding portion of the genome. Elongation of this strand continues approximately two-thirds of the way around the circular molecule, displacing the second strand, before another origin is exposed and can form a secondary structure to prime replication of the second strand (Clayton 1992). Mechanistic explanations related to this asymmetric replication process have been favored because of the distinct strand-specific distribution (Asakawa et al. 1991; Perna and Kocher 1995b) of nucleotides in animal mitochondrial genomes.

There are at least three ways that asymmetric replication could lead to compositional differences between strands of the mitochondrial genome. The structure and misincorporation patterns of replication complexes may differ between the two strands, leading to an independent equilibrium base composition for each strand. A similar compositional skew could arise if only the first strand is completely replicated on a regular basis, even if the pattern of mutation was identical for both strands (Asakawa et al. 1991). Alternatively, the transient displacement of the second strand, coupled with the difference between the mutational spectra of double- and single-stranded DNA (Lindahl 1993) could be responsible for the compositional variation between strands (Brown and Simpson 1982). This hypothesis also predicts a compositional gradient within a strand correlated with the amount of time any given position remains single-stranded. If the pattern of

mutation is governed by the availability of free nucleotides, as has been suggested for mammalian germ cells (Wolfe et al. 1989), variation in nucleotide pools during the course of one round of replication could result in strand-specific patterns as well as an intra-strand gradient. However, replication is not synchronized among genomes within a single mitochondrion or with respect to the S phase of the cell cycle (Bogenhagen and Clayton 1977), so that consistent variations in mitochondrial nucleotide pools are not expected.

Ideally, characterization of the directional mutation pressure at sites which are free from the constraints imposed by selection can provide a base-line for comparison with the pattern of evolution at other sites in the molecule. This would allow determination of the extent to which protein evolution is affected by mutational level processes and provide insight into the nature and magnitude of selective pressures acting contrary to or in concert with the mutational biases (Sueoka 1988; Sueoka 1992). Animal mitochondrial genomes are small and compact, with little or no space between coding genes except for the control region. The importance of the control region to replication and transcription of the genome suggests that its sequence will be constrained through natural selection. These constraints are reflected in the heterogeneity of base substitution rates among control regions sites (Kocher and Wilson 1991). Without an obvious neutral region for comparison, it is difficult to know whether the extreme biases at third codon positions reflect only the mutational spectrum or are also influenced by selection.

Given the wide-spread use of mitochondrial polymorphisms as neutral markers in population studies, it is perhaps surprising to note that tests of neutrality are rejected for human control region RFLP's (Merriwether et al. 1991) and *Drosophila* nucleotide sequence variation (Rand et al. 1994), even at synonymous sites (Ballard and Kreitman 1994). Natural selection could act on third codon positions in a manner unrelated to the function of the gene product. For example, acceptable nucleotides at particular third codon positions may be constrained by maintenance of secondary structure of the DNA or RNA,

or of cryptic regulatory signals necessary for replication, transcription, or post-transcriptional processing. Under these conditions, selection is not expected to lead to consistent compositional variation among codon families. In contrast, natural selection could discriminate between different synonymous codons based on the accuracy or efficiency of translation from mRNA to a functional peptide product. If translational level selection is acting on mtDNA, codon families will exhibit variation in degenerate site composition dependent on the relative fitness of synonymous codons for a particular amino acid.

Asakawa et al. (1991) provide two arguments against translational level selection in mtDNA. In some systems where translational level selection has been characterized, synonymous codon usage reflects the optimization of the level of gene expression by matching codon usage to the abundance of iso-accepting tRNAs. In mammalian mitochondrial genomes, there is only one tRNA for each degenerate codon family, except for leucine codons (*CTN* and *TTR*) which differ at the first position and have separate tRNAs. However, as Bulmer (1991) has pointed out, translational level selection does not require tRNA abundance differences and could simply result from differences in synonymous codon-anticodon binding affinity. The second argument against translational level selection in mtDNA arises from comparison of the synonymous codon usage of each strand. The frequency of synonymous codons on each strand is correlated with the overall strand-specific base composition. An inversion including a protein-coding gene in echinoderms demonstrates that synonymous codon usage evolves to match the strand-specific biases rather than retaining the original synonymous site composition (Asakawa et al. 1991). This can not be used as evidence against selection on mtDNA, because synonymous sites experiencing both a strand-specific mutational pressure and translational level natural selection may still exhibit such correlations.

Simple X^2 tests of homogeneity are one way to investigate intramolecular

compositional variation. However, this approach does not allow simultaneous consideration of the position-specific effects predicted under some mutational models and the variation among codon families that might arise from translational level selection. Subdividing complete genome data is one alternative approach that can eliminate the confounding effects of one phenomenon while allowing inference on the other. For example, we could study compositional variation around the genome by restricting the analysis to valine codons. Two obvious disadvantages of this method are a severe reduction in individual sample size and an increase in the number of individual tests. The loglinear analysis presented here allows us to make inferences about both compositional variability around mitochondrial genomes and among codon families, as well as considering these effects across several mammalian taxa with a conserved genome structure and mechanism of replication.

Methods

Categorical Variables

The data can be viewed as a multi-way contingency table, with each fourfold degenerate codon position cross-classified by SPECIES (S), CODON FAMILY (C), DISTANCE or position in the molecule (D) and nucleotide BASE (B). The value, x_{ijkl} , in each cell of the table is the number of fourfold degenerate positions of codon family j , occupied by base l , in region k of the genome of species i . Data from four species, cow (Anderson et al. 1982), fin whale (Arnason et al. 1991), harbor seal (Arnason and Johnsson 1992) and human (Anderson et al. 1981), are included in this analysis. There are eight fourfold degenerate codon families in the mammalian mitochondrial genetic code. In order to create the categorical variable for position, the protein-coding portion of the molecule has been somewhat arbitrarily divided into six regions, of approximately 2Kb each, defined by gene boundaries. Table 3.1 shows the exact position of the boundaries

used for each of the four taxa. Base, the final categorical variable, has four possible values. The complete 4x8x6x4 table has 768 cells.

Table 3.1. Exact boundaries of the discrete position categories for the variable (D). The numbers correspond to the sequence positions of the boundaries from the cow (Anderson et al. 1982), fin whale (Arnason et al. 1991), harbor seal (Arnason and Johnsson 1992) and human (Anderson et al. 1981) complete mitochondrial genome sequences. The protein-coding genes that fall within the position classes are shown in parentheses.

Position (D)	cow	fin whale	harbor seal	human
1 (cytb)	15792...13927	15891...14029	16369...14510	16023...14146
2 (nd5)	13926...12109	14028...12208	12680...14509	14145...12337
3 (nd4, nd4L)	12108...10239	12207...10339	12679...10823	12336...10470
4 (nd3, coIII, atp6, atp8)	10238...8129	10338...8228	10822...8714	10469...8366
5 (coII, coI)	8128...5687	8227...5782	8713...6275	8365...5904
6 (nd2, nd1)	5686...3101	5781...3190	6274...3680	5903...3307

Log-linear Models

If the variable BASE is used to define the columns of the contingency table, then each row of the table can be described by an independent multinomial distribution. This corresponds to a product-multinomial sampling design that is appropriate for log-linear analysis (Feinberg 1980). The logarithm of the expected cell frequencies, m_{ijkl} , can be expressed as a linear combination of terms defined by the categorical variables chosen for analysis. A fully saturated model for this data table is:

$$\log m_{ijkl} = \mu + \mu_{S(i)} + \mu_{C(j)} + \mu_{D(k)} + \mu_{B(l)} + \mu_{SC(ij)} + \mu_{SD(ik)} + \mu_{SB(il)} + \mu_{CD(jk)} + \mu_{CB(jl)} + \mu_{DB(kl)} + \mu_{SCD(ijk)} + \mu_{SCB(ijl)} + \mu_{SDB(ikl)} + \mu_{CDB(jkl)} + \mu_{SCDB(ijkl)} \quad (1)$$

where μ is the log of the total number of observations in the table, $\mu_{S(i)}$ is the effect of the

ith species, $\mu_{SC(ij)}$ is the effect of the interaction of the *ith* species and the *jth* codon family, and so on (Caswell 1989).

The parameters of the log-linear model are estimated by maximum likelihood methods using the CATMOD procedure of SAS v6.0 for Vax systems (SAS Institute, Inc., Cary, NC). The overall goodness-of-fit of the model is then assessed by the log likelihood ratio, G^2 , which is asymptotically distributed as X^2 with degrees of freedom equal to the difference between the number of cells in the table and the number of parameters estimated. The significance of individual terms is evaluated by examining the change in G^2 following the addition or deletion of that term from the model (Caswell 1989; Christensen 1990).

Only hierarchical models are considered, meaning that if a higher-order interaction term is included, all lower-order terms involving the variables in that interaction must also be included in the model. There are 113 possible hierarchical models for four-way cross-classified data if all marginal totals are free to vary. The product multinomial sampling design, however, constrains some marginal totals. In the present analysis, the three-way interaction term, $\mu_{SCD(ijk)}$, is fixed and included in all models along with the associated lower order terms, $\mu_{SC(ij)}$, $\mu_{SD(ik)}$ and $\mu_{CD(jk)}$. This reduces the total number to 19 and we can test all possible models. This is an ideal solution to the potential biases of particular model selection procedures (Feinberg 1980) and allows an investigation of which parameters are absolutely necessary in order for a model to adequately describe the data.

The notation used in equation (1) is cumbersome. The restriction to hierarchical construction allows models to be unambiguously specified by the highest order interaction term containing each variable. The saturated model shown above is [SCDB]. This notation will be used in all further references to individual models.

The residuals of a model that fits the observed data well (according to the overall goodness-of-fit test and minimization of the higher-order interaction terms) are analyzed

using the method recommended by Christensen (1990). The chosen model is refit to a function of the residuals and predicted cell values, weighting the regression by the predicted cell values, using the GLM function of the MINITAB Statistical Package for Vax (MINITAB Inc., State College, PA). This provides the standardized residuals, leverages and Cook's D values for the log-linear model which are analogous to the standard measures used to evaluate the residuals of a linear regression analysis.

Results and Discussion

The overall base composition of fourfold degenerate sites from the four taxa (table 3.2) illustrates several well known features of mtDNA. The strand asymmetry (Asakawa et al. 1991, Saccone et al. 1993, Perna and Kocher 1995) is evident in comparisons between %G and %C and between %A and %T. All four genomes show a bias toward codons ending in A or C for genes on this strand and the human sequence uses far more codons ending in C and fewer ending in A than the other mammals (Anderson et al. 1991).

Table 3.2. Overall composition of fourfold degenerate third codon positions from 12 mitochondrial genes encoded on one strand in the cow (Anderson et al. 1982), fin whale (Arnason et al. 1991), harbor seal (Arnason and Johnsson 1992) and human (Anderson et al. 1981) genomes.

Taxon	%G	%A	%T	%C
cow	4.38	51.20	15.83	28.59
fin whale	2.84	48.74	16.10	32.31
harbor seal	7.39	49.97	13.18	29.46
human	4.95	39.45	13.41	42.19

Table 3.3 shows the 19 possible loglinear models, associated G^2 and degrees of freedom. Eight models were not fit to the complete data set because the overall

compositional bias against *G* in mitochondrial genomes leads to excessive zeros. The overall goodness-of-fit is rejected for 10 of the remaining 11 models at the $p=0.001$ level. Only the model [SCD][SB][CB][DB] provides acceptable fit to the complete data set. In addition to the terms fixed by design, this model includes the three second-order interaction terms involving BASE. Under this model, the multinomial distribution describing base composition is not independent of species, codon family or position in the molecule.

A subset of the data, excluding the three least frequent codon families (alanine, arginine, and proline), was used to investigate models which could not be fit to the complete data set. Seven of the eight models provide an acceptable fit to the five codon family data (table 3.3). No model rejected in the complete data analysis is accepted for the reduced data set. The simplest model that fits the reduced data is the same model that fit the complete data. This model suggests that the familiar pattern of base composition seen in table 3.2 reflects an average of data collapsed over heterogeneous regions of the genome and heterogeneous codon families.

Table 3.3. Nineteen possible log-linear models. $\sqrt{}$ marks models that have a conditional independence interpretation. \emptyset indicates that the model was not fit to the data because of zeros in the marginal values.

MODEL	8 codon families		5 codon families	
	G^2	df	G^2	df
[CSDB]	fully saturated model			
[SCD][SCB][SDB][CDB]	\emptyset		196.42	180
$\sqrt{}$ [SCD][SCB][SDB]	\emptyset		257.42	240
$\sqrt{}$ [SCD][SCB][CDB]	\emptyset		250.31	225
$\sqrt{}$ [SCD][SDB][CDB]	\emptyset		243.12	216
[SCD][SCB][DB]	\emptyset		311.17	285
[SCD][SDB][CB]	\emptyset		302.54	276
[SCD][CDB][SB]	\emptyset		297.70	261
$\sqrt{}$ [SCD][SCB]	\emptyset		390.73**	300
$\sqrt{}$ [SCD][SDB]	886.69***	504	453.08***	288
$\sqrt{}$ [SCD][CDB]	601.87***	436	403.83***	270
[SCD][SB][CB][DB]	572.73	528	356.95	321
$\sqrt{}$ [SCD][SB][CB]	688.23***	543	436.39**	336
$\sqrt{}$ [SCD][SB][DB]	930.73***	549	506.24***	333
$\sqrt{}$ [SCD][CB][DB]	724.39***	537	462.96***	330
$\sqrt{}$ [SCD][SB]	1039.72***	564	581.92***	348
$\sqrt{}$ [SCD][CB]	839.72***	552	542.73***	345
$\sqrt{}$ [SCD][DB]	1076.89***	558	608.03***	342
$\sqrt{}$ [SCD][B]	1186.34***	573	684.27***	357

unmarked G^2 values are associated with $p > 0.05$

** G^2 values significant at the $p < 0.001$ level

*** G^2 values significant at the $p < 0.0001$ level

Residual Analysis

The residuals of this simplest model, [SCD][SB][CB][DB], were analyzed to detect observations that were especially influential and to validate the assumptions of the model. A simple boxplot of the standardized residuals (figure 3.1) reveals five mild outliers out of the 768 residuals corresponding to the cells of the complete data table ($5/768=0.65\%$). This is consistent with the expectations of a standard normal distribution.

Leverages are frequently used in standard regression analysis to measure the magnitude of deviations of a particular cell value from the average of all cells. The selected model has 528 *df* and there are 768 cells, so the sum of the leverages adds up to $768-528=240$. The average leverage is $240/768=0.3125$ and 29 of the 768 leverages exceed twice the average (figure 3.2). However, no leverage is greater than 0.7183. Although these leverages are not exceptionally high, all 29 cells involve the base adenine and leucine (15), arginine (12), or occasionally valine (2) codons. Since the [SCD][SB][CB][DB] model is chosen with or without the inclusion of the arginine codon family, it is unlikely that these cells are exerting undue influence on the fit of the model.

Cook's distances measure the influence a particular observation has on the fit of the model by dropping the corresponding cell and refitting the model. Although the statistic is best suited for a Poisson sampling scheme, it can be adapted to product multinomial sampling by reducing the *df* by the number of independent multinomials. The largest Cook's distance for this data and model is 0.0416, while the critical value for substantial influence is 0.9861, indicating that no one cell has an excessive influence on the fit of this model.

Figure 3.1. Boxplot of the standardized residuals. The five outliers correspond to the cells containing observations from distance class 1, fin whale glycine codons ending in the base A, st. res=-2.805, obs=6, exp=12.6137; distance 2, human alanine codons ending in A, 2.806, 20, 12.0448; distance 2, human, valine, T, 3.246, 6, 1.9008; distance 6, fin whale, arginine, G, 3.077, 0, 0.4126; and distance 6, fin whale, threonine, C, -2.873, 14, 24.7788.

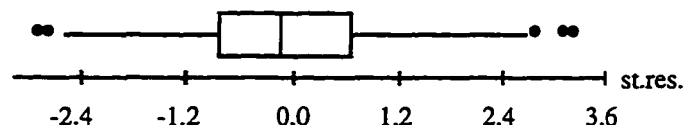
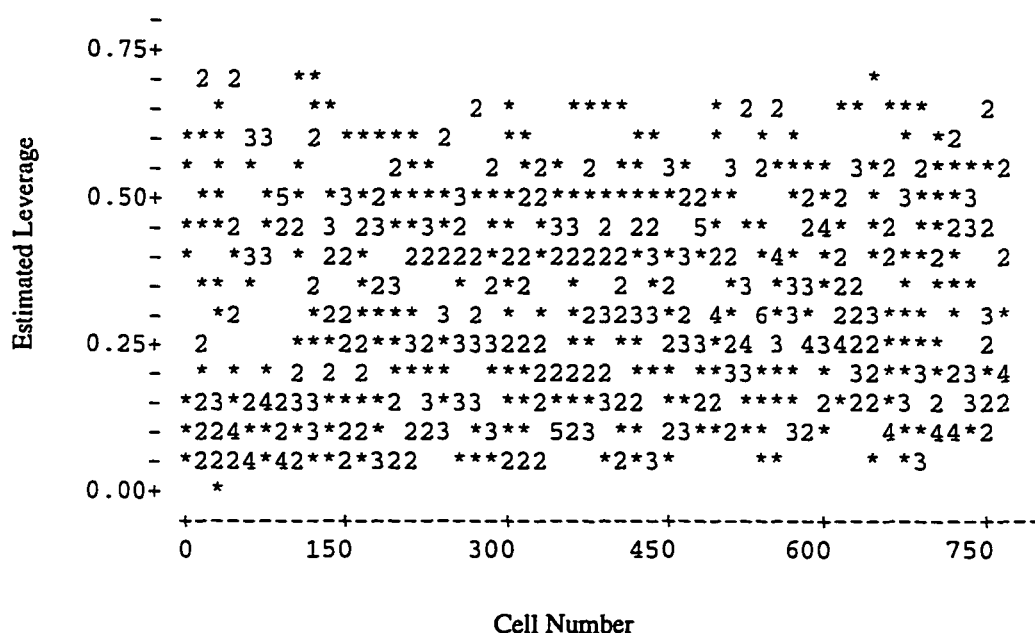


Figure 3.2. Estimated leverages vs. cell number for 768 cells of the data table. The mean leverage is 0.3125. Numbers indicate multiple observations of leverages indistinguishable at the resolution of this plot.



Conditional Independence Models and Evolutionary Hypotheses

Although the model [SCD][SB][CB][DB] appears to fit the data quite well, several conditional independence models (marked with a \checkmark in table 3.3) are of interest for examining hypotheses about the evolutionary origin of nonrandom synonymous codon usage. Hypotheses about directional mutation pressures in mtDNA can be divided into two categories: those which predict a uniform distribution of each base on a given strand and those which predict an intramolecular gradient of each base. Adequate fit of models where BASE is independent of DISTANCE, given CODON FAMILY and/or SPECIES would support the former category of hypotheses.

No models involving conditional independence of BASE and DISTANCE fit the complete or reduced data. Table 3.4 illustrates the significance of the [DB] interaction term indicating that this analysis does not support a uniform base composition around the mammalian mitochondrial genomes even when we take into consideration compositional differences among codon families, [CB] or [SCB]; the variation in codon usage around the molecule, [CD]; and species level differences in composition, [SB] or [SCB]. The level of significance of the [DB] term remains constant whether or not a higher-order interaction term [SCB] is included.

Similarly, adequate fit of models in which BASE and CODON FAMILY are conditionally independent would be consistent with Asakawa et al.'s (1991) prediction that translational level selection is not an important factor in mtDNA base substitution. The G^2 values in table 3.2 show that no model adequately fits the data unless it contains the [CB] term. Table 3.5 shows the significance of the [CB] term evaluated by the change in G^2 which occurs when it is removed from a model. Synonymous codon usage clearly differs among fourfold degenerate codon families.

Table 3.4. The significance of [DB] term in log-linear models of the 5 codon family data.

Model	G^2	df	p
[SCD][SB][CB]	436.39	336	
[SCD][SB][CB][DB]	356.95	321	
[DB]	79.44	15	$p<0.001$
[SCD][SCB]	390.73	300	
[SCD][SCB][DB]	311.17	285	
[DB]	79.56	15	$p<0.001$
[SCD][B]	684.27	357	
[SCD][DB]	608.03	342	
[DB]	76.24	15	$p<0.001$

Table 3.5. The significance of [CB] term in log-linear models of the 5 codon family data.

Model	G^2	df	p
[SCD][SB][DB]	506.24	333	
[SCD][SB][CB][DB]	356.95	321	
[CB]	149.29	12	$p<0.001$
[SCD][SDB]	453.08	288	
[SCD][SDB][CB]	302.54	276	
[CB]	150.54	12	$p<0.001$
[SCD][B]	684.27	357	
[SCD][CB]	542.73	345	
[CB]	141.54	12	$p<0.001$

The necessity of including [SB] in an acceptable model, demonstrated in table 3.6, is perhaps less surprising given the considerable literature indicating that human mtDNA base composition is distinct among mammals (Anderson et al. 1981; Lanave et al. 1984). In comparison, compositional differences among the three remaining taxa, the cow, fin whale and harbor seal appear small; However, the [SB] term remains significant when models are refit to a data set excluding the human genome (data not shown).

Table 3.6. The significance of [SB] term in log-linear models of the 5 codon family data.

Model	G^2	df	p
[SCD][CB][DB]	462.96	330	
[SCD][SB][CB][DB]	356.95	321	
[SB]	106.01	9	$p<0.001$
[SCD][CDB]	403.83	270	
[SCD][CDB][SB]	297.70	261	
[SB]	133.13	9	$p<0.001$
[SCD][B]	684.27	357	
[SCD][SB]	581.92	348	
[SB]	102.35	9	$p<0.001$

Since the fourfold degenerate sites from the four taxa may not be independent because of shared evolutionary history, it is reasonable to question whether the necessity of including [CB] and [DB] interaction terms is an artifact of assuming independence. Data from each of the four mammals constitutes a three-way contingency table and the significance of the [CB] and [DB] terms can be evaluated for each species individually. We fit four models to each subset of the five codon family data. The results are shown in table 3.7. The model which includes all two-way interaction terms [CD], [CB] and [DB] fits the

data from each of the four mammals quite well. The model in which BASE is conditionally independent of CODON FAMILY given DISTANCE, [CD][DB], is rejected for all four mammals indicating that the synonymous codon usage differs among codon families within all four mitochondrial genomes. The model in which BASE is independent of all other variables, [CD][B], is uniformly rejected. The model in which BASE is conditionally independent of DISTANCE given CODON FAMILY, [CD][CB], is rejected for both the fin whale and the harbor seal but provides an adequate fit to the data from human and cow mtDNA. However, the Akaike Information Criterion (AIC) (Christensen 1990) clearly favors the [CD][CB][DB] model over either conditional independence model for all four taxa, supporting the idea that inclusion of a [DB] term provides significantly more information about the data.

Table 3.7. Log-linear models to test the significance of [DB] and [CB] terms for each mammal individually. * indicates that the overall G^2 is significant at the 0.05 level.

Model	<i>df</i>	human		cow		fin		seal	
		G^2	AIC	G^2	AIC	G^2	AIC	G^2	AIC
[CD][DB][CB]	60	61.05	61.05	52.34	52.34	73.11	73.11	70.92	70.92
[CD][DB]	72	120.25*	144.25	98.93*	122.93	111.26*	135.26	122.64*	146.64
[CD][CB]	75	83.73	113.79	81.49	111.49	99.66*	129.66	125.85*	155.85
[CD][B]	87	141.31*	195.31	127.11*	181.11	133.94*	187.94	179.56*	233.56

Base Composition and Position in the Genome

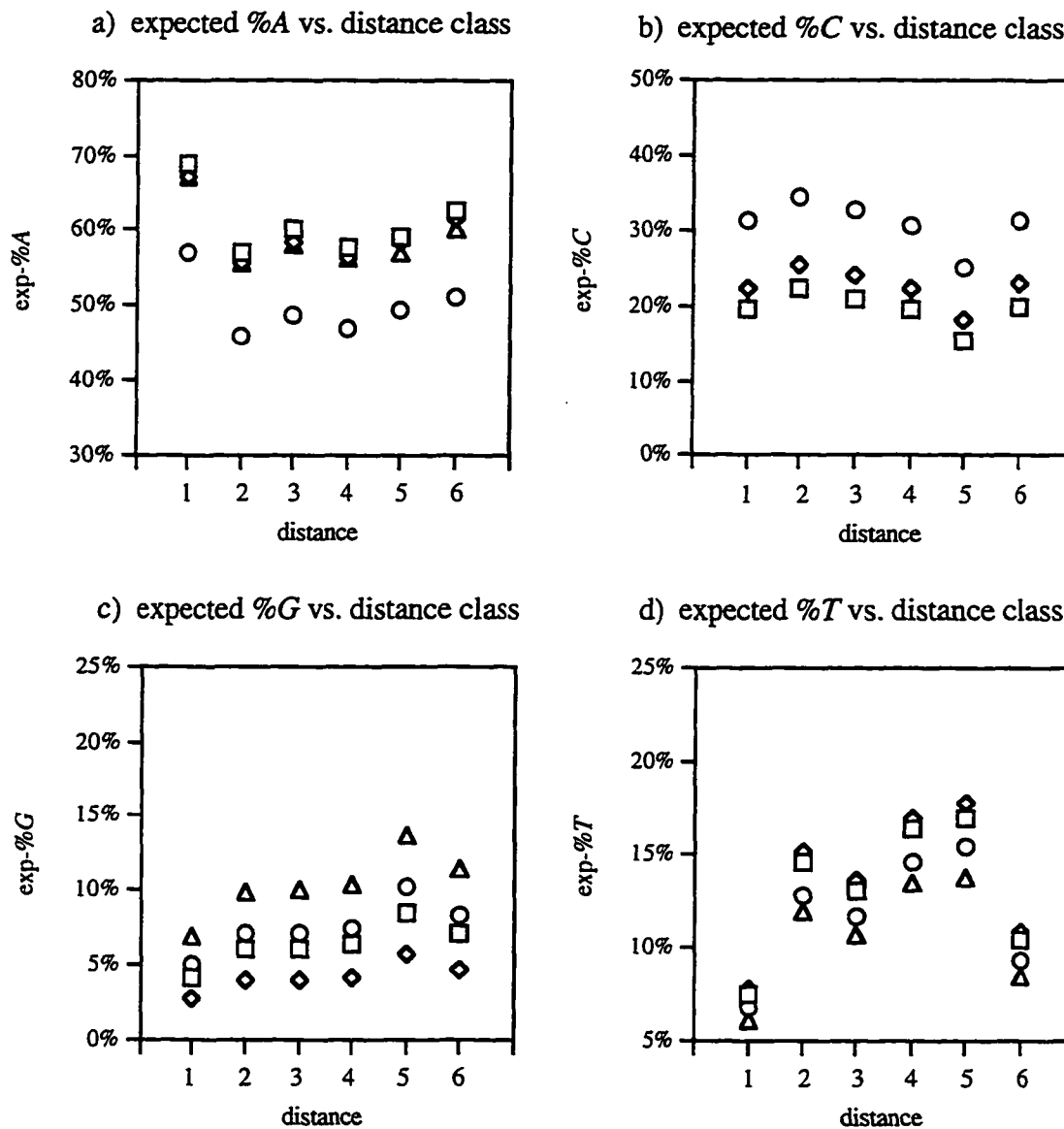
The simple observation that composition varies among the six defined regions of the mitochondrial genome is not sufficient evidence to conclude that there is a compositional gradient around the molecule. Expected values for the 768 cells of the complete data set given the [SCD][SB][CB][DB] model can be used to illustrate the predicted relationship between BASE and DISTANCE. Figure 3.3 shows this relationship

using data from leucine codons. Under the constraints of this simple model, the relationship between distance and the frequency of any given base is the same for all codon families although the expected frequency of the base differs among codon families. Similarly, it is obvious from figure 3.3 that the predicted relationship is constant across species although the human genome exhibits a lower relative frequency of *A* and a higher relative frequency of *C* than the other three mammals.

The predicted variation in base frequency with position in the molecule in figure 3.3 provides some support for the idea of gradual compositional changes around the molecule. The relative frequency of *A* is highest in distance class 1, then falls to its minimum in distance class 2 and rises gradually in the remaining distance class levels. The relative frequency of *C* diminishes from distance class 1 to distance class 5 then rises abruptly in distance class 6. The opposite pattern is shown by both *G* and *T* which are at a maximum in distance class 5 then fall in class 6. This is of interest since the second origin of replication for mammalian mitochondrial genomes lies between distance classes 5 and 6.

Directional mutation pressures arising from differences between double- and single-stranded DNA mutational spectra predict compositional gradients correlated with time spent single-stranded. The replication mechanism involves displacement of one strand, often referred to as the heavy strand, beginning at the first origin of replication in the D-loop immediately preceding distance class 1. Each successive distance class defined in this analysis will spend a smaller proportion of time single-stranded than the preceding distance class except distance class 6 which occurs after the second origin of replication and therefore remains single stranded considerably longer. W. K. Thomas (personal communication) has predicted that the frequency of *G* on the strand considered in this analysis will show an inverse relationship with time spent single-stranded because of the difference in the frequency of cytosine deamination between double and single-stranded DNA. Figure 3.3 does in fact support a relationship consistent with this prediction.

Figure 3.3. Predicted percent of the four nucleotides in each distance class for the cow (squares), fin whale (diamonds), harbor seal (triangles), and human (circles) fourfold degenerate third codon positions. Predicted values are based on the [SCD][SB][CB][DB] model.



Although the simple model used to generate the predicted values for figure 3.3 fit the data well, any model containing higher-order interaction terms in addition to all six two-way interaction terms is also statistically acceptable. For instance, the model [SCD][SCB][SDB][CDB], that contains all possible terms except the four-way interaction, is favored over the simple model on which figure 3.3 is based, when evaluated using the *AIC*. These models will be discussed in a later section of this paper, but it is important to realize that the relationship between base composition and distance class shown in figure 3.3 represents only one of several alternative patterns arising from this log-linear analysis. Interpretation of the predicted relationship between base composition and distance for the higher-order models is not trivial. The addition of a single three-way interaction term can lead to predicted relationships which vary considerably between species and among codon families from a single species. Although there is no simple qualitative description which encompasses all predicted relationships from all higher-order models, it is safe to say that some higher-order models do not predict a relationship that can be interpreted as support for the idea that compositional gradients exist in mammalian mitochondrial genomes.

Base Composition and Codon Family

Translational level selection can exert a directional pressure on synonymous codon usage by increasing the proportion of “optimal” codons in highly expressed genes. “Nonoptimal” codon usage arises in genes with a lower expression level due to less selective pressure and an equilibrium between selection, mutation and genetic drift rather than from selection for a reduced level of expression (Bulmer 1991). Differences in synonymous codon usage among codon families can arise from codon family specific codon-anticodon interaction dynamics. The observation that synonymous codon usage differs among the eight fourfold degenerate codon families included in this analysis (and even among the five codon families in the reduced data set) suggests that we should not

dismiss the possibility that translational level selection may be acting in mammalian mtDNA.

Traditional methods for documenting translational level selection do not work well for mtDNA. All 13 protein-coding genes in mammalian mtDNA encode important components of oxidative phosphorylation machinery including subunits of the ATP synthetase, NADH dehydrogenase, the *b-c₁* and cytochrome oxidase complexes. If there is no clear differential expression, we cannot provide evidence of translational selection by comparing “optimal” usage among highly expressed genes and “nonoptimal” usage among other genes. In fact, any comparisons among genes are difficult because of the extreme economy of the mitochondrial genome. The largest mammalian mitochondrial protein, ND5, is about 600 amino acids long and 4 of the 13 mitochondrial proteins have less than 200 amino acids each. Considerable variance in codon usage among genes may result from small sample sizes from individual genes. Even if we disregard this source of error, 13 is a very small total number of genes especially compared to sample sizes typically used to investigate translational level selection by correspondence analysis. Furthermore, the absence or a reduced level of translational selection is often inferred by demonstrating the equality of complementary dinucleotide frequencies (Shields and Sharp 1987). This test is inappropriate for mtDNA because of the strand-specific nature of mitochondrial directional mutation pressures.

We might instead consider translational level selection acting on the 13 mitochondrial protein-coding genes as a unit and postulate that optimal codon usage is common to all genes. Variation in codon usage among these genes might still arise from position-specific directional mutation pressures, but variation among codon families at a particular location in the molecule would result from codon family-specific codon-anticodon affinities. A single anticodon must recognize all four codons of a fourfold degenerate codon family. These mitochondrial anticodons all have the same base, *U*, in the

first position (Anderson et al. 1981; Anderson et al. 1982; Arnason et al. 1991; Arnason and Johnsson 1992). We might expect that if there is a unique “optimal” codon for each codon family (Bulmer 1991), it would be the codon ending with *A*, for all eight fourfold degenerate families, since this codon would form a natural Watson-Crick basepair with the anticodon. If we examine each codon family, at each distance, from each species individually (for a total of 192 samples of synonymous codon usage) we find 128 instances where *A* is the most frequent third base of the codon. Predicted values for [SCD][SB][CB][DB] show 148 out of 192 instances where *A* is the most frequent third base. The predicted highest frequency codon always ends in *C* if it does not end in *A*. Even in human samples where *C* is the most frequent third base overall, some codon families still use codons ending in *A* more often than those ending in *C*. For example, *CTA* leucine codons are always more frequent than *CTC* codons. In taxa other than the human, the only codon families that use *C* as the most frequent third base are alanine and proline.

We might also expect that the intensity of selection for optimal usage is dependent on the structure of the codon. For example, codons that form *GC* base pairs with the anticodon in the first two codon positions may differ in binding affinity from codons which form only *AT* base pairs in these positions. Four of the codon families considered in this analysis will form two *GC* base pairs with the anticodon in codon positions 1 and 2 (ala *GCN*, gly *GGN*, pro *CCN* and arg *CGN*). The remaining four codon families have one *GC* and one *AT* base pair in these positions. These four can be distinguished into two groups according to whether the first position (leu *CTN* and val *GTN*) or the second position (ser *TCN* and thr *ACN*) is involved in the *GC* base pair with the anticodon. If the data set is restricted to the codon families within a structural group, we can test again for conditional independence of *BASE* and *CODON FAMILY* given *SPECIES* and *DISTANCE* to determine whether codons with similar structure have similar synonymous

codon usage patterns. Both relevant models [SCD][SB][DB] and [SCD][SDB] fit the data set composed of leucine and valine codons ($G^2=139.68$, $df=117$, $p=0.0750$ and $G^2=90.01$, $df=92$, $p=0.0741$). The [SCD][SDB] model cannot be fit to the data set composed of serine and threonine codons, but the simpler conditional independence model fits the data quite well ($G^2=76.54$, $df=117$, $p=0.3391$). For the remaining structural group data set, composed of alanine, glycine, proline and arginine codons, both conditional independence models are rejected at the $p=0.05$ level, indicating compositional heterogeneity. If this latter group is subdivided by the base at the second codon position, we can clearly demonstrate compositional homogeneity of alanine and proline ($G^2=114.87$, $df=117$, $p=0.5385$), but still reject the simple conditional independence model for the group composed of glycine and arginine codons ($G^2=146.91$, $df=117$, $p=0.0320$). Third codon position nucleotide usage may be related to the stability of codon-anticodon interactions. Selection favors the formation of a natural Watson-Crick base-pair with the wobble position of the anticodon, but the strength of selection on third base usage depends on codon-anticodon interactions at the first two codon positions.

Higher-order Interactions

The predicted relationships among the variables in this analysis are dependent on which, if any, higher-order interaction terms are included in a model. There are several quantitative means of evaluating the effect of including these terms in a model (Christensen 1990). One relevant measure is the adjusted R^2 ; Another is the AIC . Both methods favor any and all higher-order models over the simplest model that is acceptable based on the G^2 . There are reasonable biological interpretations for higher-order terms, such as [SCB], which could be indicative of differences in translational level selection among species. A [SDB] term may indicate that compositional variation around the molecule arises in a taxon specific manner, perhaps due to differences among species in the DNA polymerase

misincorporation spectra for ssDNA and dsDNA. The [SCD][SB][CB][DB] model explains nearly 60% of the total variation in this data ($R^2=0.5965$). Inclusion of a single three-way interaction term increases this to 71.32-75.57%. Addition of a second three-way interaction term improves the explanatory power to 80.72-83.45%, and the model which includes all three three-way interaction terms explains 88.23% of the total variation.

Dinucleotide and Higher-order Mutational Biases

Patterns of mutation at a particular nucleotide site can be influenced by context (Bulmer 1990). The reduced frequency of *CpG* dinucleotides in *E.coli* and yeast (Bulmer 1990), presumed to result from increased mutability conferred by methylation, is perhaps the best known example of a mutational bias influenced by adjacent base. While there is no known methylation activity in animal mtDNA, it is quite possible that there are some context specific mutational mechanisms acting in this system. These contextual biases are not limited to the bases immediately preceding and following the target site, but could extend, at least theoretically, to trinucleotides, tetranucleotides or farther. Higher-order mutational biases could be responsible for the observed differences in synonymous codon usage among fourfold degenerate codon families. Such biases might also contribute to the remaining variation in this data which cannot be accounted for by the parameters investigated in this analysis.

Eyre-Walker (1991) used a chi-square independence test to compare third codon position nucleotide frequencies of the four fourfold degenerate codon families that have *C* at the second position. This *C-test* can be performed on data resampled to correct for the frequency of first base of the following codon, which should eliminate the effects of mutational biases at the dinucleotide level. Unfortunately, the reduction in sample size resulting from resampling makes it impossible to use the *C-test* for individual distance classes as defined in our log-linear analysis. Collapsing the data over all distance classes

may be unwise since this analysis supports the idea that directional mutation pressures in mammalian mitochondrial DNA vary around the genome. Furthermore, the predictions of a translational level selection model for mitochondrial DNA do not exclude the possibility that codons which have the same second position base have similar codon-anticodon interactions and hence similar third codon position nucleotide usage. Keeping these potential limitations in mind, we have applied the *C-test* to the same data used in this analysis, pooled across distance classes. The *C-test* does not detect any significant differences ($p < 0.05$) among serine, proline, threonine and alanine third codon position base usage in either the cow or seal genome, with or without resampling to equalize the frequency of the 'fourth' position. Differences among these codon families are significant in both the fin whale and human mitochondrial genomes and the magnitude and the significance of these differences are largely unaffected by the resampling procedure. Hence, nearest neighbor effects cannot completely explain the compositional heterogeneity among codon families.

Other sources of Selection on Synonymous Sites

There are several additional potential sources of selective pressure that could influence nucleotide usage at synonymous sites in mtDNA and might explain some of the remaining variation in this data. Delorme and Henaut (1991) have suggested that base composition of mitochondrial DNA varies predictably with position in the polycistronic mRNA transcript and may be important for the regulation of RNA polymerase activity. This cannot be ruled out as an alternative explanation for the interaction we observe between distance class and base composition. Selection might act to preserve or prevent secondary structure in either the mRNA transcript or the ssDNA replication intermediate. Selective pressures which differ among mitochondrial genes could also lead to compositional differences among distance classes. Post-transcriptional mRNA processing

does occur in mtDNA (Clayton 1992) and cryptic signals for cleavage of the transcripts and polyadenylation have not been characterized. There is little or no 5' untranslated sequence for any fully processed mitochondrial transcript and the mechanism by which ribosome binding and translation initiation occur is uncertain, but this might also provide a selective constraint on synonymous site base composition.

Conclusions

The general conclusions that base composition varies among species, codon families and regions of the circular genome are strongly supported by this analysis since all models providing an acceptable fit to the data include all two-way interaction terms involving BASE. There is tentative support for compositional gradients predicted by a model of directional mutation pressure based on the asymmetric replication and the difference between mutational spectra of double- and single-stranded DNA. Translational selection may be responsible for compositional differences among codon families with dissimilar codon structure. This may complicate evaluation of the magnitude of directional mutation pressures in mtDNA if it is necessary to assume that the extreme compositional patterns observed at fourfold degenerate sites represent equilibrium frequencies of the four bases generated solely by a mutational process.

CHAPTER IV

INTRAMOLECULAR PATTERNS OF SYNONYMOUS BASE COMPOSITION IN TWO ADDITIONAL TAXONOMIC GROUPS: INSECTS AND MOLLUSCS

Base composition of synonymous codon positions varies among animal mitochondrial genomes and between strands of single mitochondrial genomes (Perna and Kocher 1995b). Log-linear analyses of four mammalian mitochondrial genomes revealed that base composition also varies within a strand depending on the codon family and position in the molecule (Chapter III). These intramolecular patterns of base composition provide clues about the evolutionary forces that shape mitochondrial genomes. However, the compositional patterns reflect a complex balance of phenomena, and it is not simple to separate the effects of non-random mutational biases and natural selection. This analysis explores intramolecular patterns of compositional variation in two additional taxonomic groups to complement the studies of mammalian mitochondrial genomes.

Introduction

Log-linear analysis of four mammalian taxa (Chapter III) indicates that base composition is not independent of species, position in the molecule or codon family. Predicted values for a simple model provide tentative evidence of a compositional gradient around the molecule compatible with predictions of a directional mutation pressure related to time spent single-stranded during replication. The lack of independence of base and codon family suggest the action of either translational level selection or dinucleotide mutational biases. Simple examination of mammalian base composition with knowledge of the Watson-Crick base-pairing rules reveals that composition also varies between strands of these genomes,

as quantified in Chapter II (Perna and Kocher 1995). In mammalian mitochondrial genomes, 12 of the 13 protein coding genes are encoded on the same strand. The single remaining gene, ND6, does not contain a sufficient number of fourfold degenerate sites to allow inclusion of strand as an additional categorical variable in the log-linear analysis. The strand asymmetry is accommodated by always including a term to describe the deviations of a given base from the mean of all four bases and restricting the data to fourfold degenerate sites from genes on the major coding strand.

Genes are more evenly distributed between the two strands of mitochondrial genomes from two molluscs, a black chiton, *Katharina tunicata* (Boore and Brown 1994) and a pulmonate gastropod, *Cepaea nemoralis* (Terrett et al. 1995). Three insect genomes, from a fruit fly, *Drosophila yakuba* (Clary and Wolstenholme 1985); a mosquito, *Anopheles gambiae* (Beard et al. 1993); and a honeybee, *Apis mellifera* (Crozier and Crozier 1993), also encode multiple peptides on both strands. With minor modifications, the log-linear analysis presented in Chapter III can be adapted to examine compositional patterns among codon families and around these genomes while simultaneously considering the variation between strands. Analysis of these additional genomes will establish if the observed intramolecular patterns are restricted to mammalian genomes and should provide additional insight into the evolutionary mechanisms generating mitochondrial base compositional biases.

If mutational biases alone shape the base composition of fourfold degenerate sites and the pattern of mutation is identical for the two strands of the genome, then sampling the composition of different sites on each strand will provide two different point estimates of the same multinomial distribution. Furthermore, Sueoka (1995) has demonstrated that equal frequency of A and T and of G and C within each strand is a logical consequence of Watson-Crick base-pairing rules under these conditions.

If, as appears to be the case for most metazoan mitochondrial genomes, the pattern of

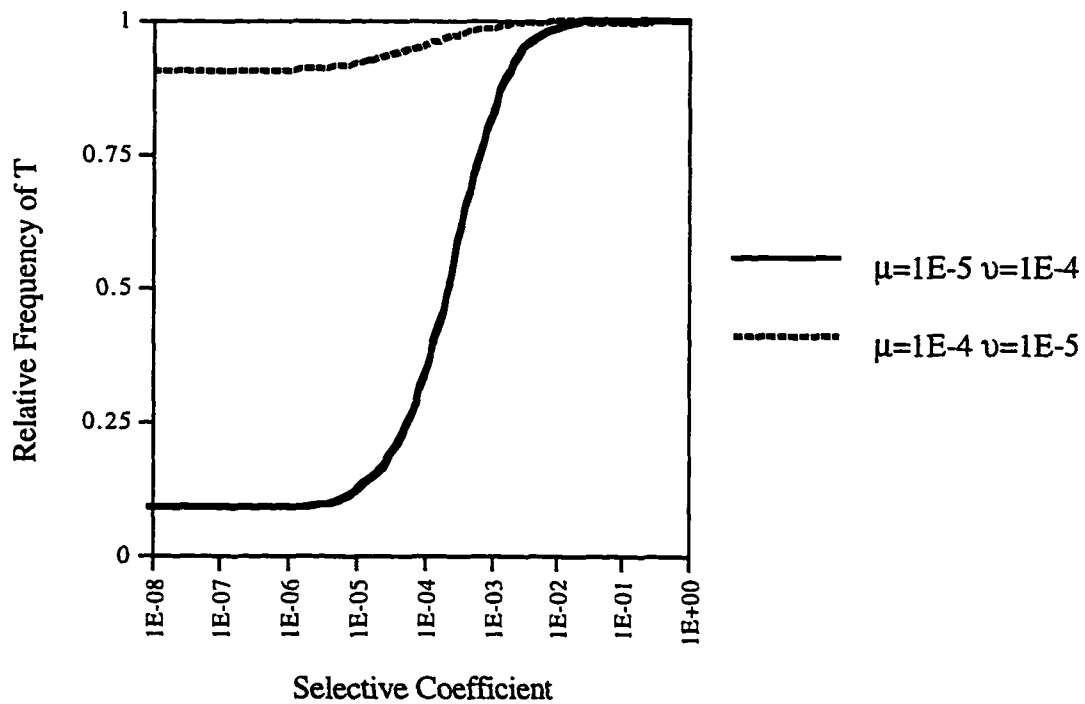
mutation varies between strands, the base composition of one strand will not be identical to the composition of the other. Instead, the composition of the first strand will be a mirror image of the composition of the second strand, such that the frequency of *G* on the first is equal to the frequency of *C* on the second, and so on. Under these conditions, the equality of *A* and *T* (*G* and *C*) within a strand is no longer guaranteed and the sequence may exhibit skew (Chapter II). Sampling fourfold degenerate sites from each strand should provide estimates of a multinomial distribution and its 'reverse complement'.

If synonymous codon usage is entirely determined by natural selection for optimal translation, then fourfold degenerate site composition for each codon family will be identical for both strands. Composition of third positions can vary among codon families as a result of differences in the relative fitness of synonymous codons for a particular amino acid. Under this model, the base composition of fourfold degenerate sites from each strand would reflect a heterogeneous mixture of eight different multinomial distributions. The composition of the two strands could vary simply as a result of differences in amino acid composition, but a single multinomial distribution should describe the composition of both strands for a given codon family.

Thus, if mutational biases and translational level selection were mutually exclusive, it would be relatively straightforward to interpret compositional patterns. Other studies of mitochondrial base composition have provided strong support for the idea that mutational patterns are biased in a strand-specific manner (Asakawa et al. 1991, Tanaka and Osawa 1994). If natural selection is acting as a filter, screening out sub-optimal codons, it is not likely to be provided a random set of mutations to sort. This alters the compositional expectations in a manner difficult to predict without prior knowledge of the underlying mutational pattern and the magnitude and direction of selection. For simplicity, consider base composition of a sequence composed entirely of *AT* base pairs. Basic population genetics illustrates the expected equilibrium between mutation and selection for a two

character state model (figure 4.1). The two curves represent the two strands of a single genome where the rate of change from A to T is an order of magnitude greater than the rate of change from T to A on one strand. Base-pairing ensures that the other strand has a rate of change for T to A an order of magnitude greater than the rate from A to T. When the selective differential between T and A is small, there is no change in the base composition from the equilibrium predicted from the mutation rates. As the selective differential grows, the equilibrium composition is driven toward higher frequency of the more optimal character state. However, the equilibrium changes at different rates for the two strands.

Figure 4.1. Mutation-selection equilibrium frequency of advantageous character state for two underlying mutational pressures, where μ is the rate of mutation from A to T, and ν is the rate of change from T to A. The equilibria are estimated for a group of tightly linked sites in a large population based on the formulation found in Bulmer (1990).



A further complication for interpretations of base compositional patterns is the observation that the patterns of mutation at a particular nucleotide site can be influenced by context (Bulmer 1990). These contextual biases are not limited to the bases immediately preceding and following the target site, but could extend, at least theoretically, to trinucleotides, tetranucleotides or farther. One hydroxyl adduct of guanine formed in the presence of oxygen free radicals is known to exhibit some contextual bias and may be a significant factor in mitochondria where metabolic byproducts of oxidative respiration are found in high concentration. Tanaka and Osawa (1994) were unable to detect the contribution of this specific mutational pathway to the overall mutational matrix compiled from 43 mitochondrial genomes of diseased humans, although they attribute some heterogeneity in substitution patterns among sites to this or other context dependent mutational mechanisms. Under a dinucleotide bias model, the mutational spectrum at the third codon position of a codon family is dependent on the nucleotide at the second codon position and the composition of first positions of the following codons. Codon families that have different bases at the second codon position may exhibit compositional variation at the third position. Even codon families with the same base at the second position can vary at the third, if the distribution of adjacent codons varies because of selection for amino acid sequence of the encoded peptides. If single nucleotide mutational patterns are different for the two strands of the genome, dinucleotide biases may also vary between strands. The predictions of higher-order mutational bias models appear to be indistinguishable from the predictions of a balance between simple mutational biases and selection. The Eyre-Walker (1991) *C-test* compares third codon position nucleotide frequencies of the four fourfold degenerate codon families that have C at the second position from data resampled to correct for the frequency of first base of the following codon. Application of this test to each of the four mammalian genomes used in the log-linear analysis demonstrated that correcting for dinucleotide biases did not eliminate all compositional heterogeneity among codon

families (Chapter III). Therefore, dismissal of translational level selection as a potential contributor to intramolecular compositional patterns (Asakawa et al. 1991) seems premature.

Clearly, the evolutionary mechanisms generating base compositional variation, even at fourfold degenerate sites, in mitochondrial DNA can be complex and difficult to expose. A better description of the pattern of variation we are trying to explain is an obvious step toward a better understanding. This analysis expands our previous consideration of compositional variation among and within metazoan mitochondrial genomes, and explores whether any or all hypothesized evolutionary models are consistent with observed patterns of variation.

Methods

This log-linear analysis of five invertebrate taxa follows the basic methodology described in Chapter III for the log-linear analysis of four mammalian genomes. Here each genome is treated individually, eliminating the variable SPECIES. A new variable, STRAND (S) is added to examine intramolecular variation between the two strands of each genome. Consequently, for each taxon there are 19 possible log-linear models for the four way cross-classified data table with the marginal totals for the three-way interaction term, $\mu_{SCD(ijk)}$, and all associated lower order terms fixed by the product multinomial sampling design.

Figure 4.2 illustrates the organization, relative size and direction of transcription for the 13 protein coding genes of the *Cepaea nemoralis* mitochondrial genome (Terrett et al. 1995). The origins of replication and transcription have not been characterized for *Cepaea* mtDNA, so unlike the mammalian log-linear analyses, the distance classes can not be defined relative to these landmarks. Figure 4.2 shows the boundaries of the four arbitrary distance classes beginning immediately after one of the two largest non-coding regions.

Fourfold degenerate third codon positions from *Cepaea* are cross-classified by STRAND (S), DISTANCE CLASS (D), CODON FAMILY (C) and BASE (B), creating a $2 \times 4 \times 8 \times 4 = 256$ cell data table. Distance classes 1 and 4 contain only genes encoded on one strand. Therefore, $1 \times 2 \times 8 \times 4 = 64$ cells are fixed zeros leading to a reduction in the total degrees of freedom for the log-linear analysis. The *Katharina tunicata* genome (Boore and Brown 1994) is also partitioned into four distance classes beginning from the largest non-coding region (figure 4.3) creating a 256 cell data table with 64 fixed zeros. However, comparison of figures 4.2 and 4.3 reveals that there have been multiple rearrangements of these two molluscan mitochondrial genomes since their divergence over 500 MYA (Terrett et al. 1995). Like most metazoan genomes, these two encode the same 13 peptides; However, the gene order is quite different. In *Katharina*, 7 genes are encoded on one strand and 6 are encoded on the other, while in *Cepaea*, 9 are encoded on one strand and 4 on the other. Again, there is no knowledge of replication patterns for these two taxa and consequently, it is impossible to determine which strands are homologous with respect to this process. These differences in genome architecture may be important considerations in comparisons of the results of the two log-linear analyses of molluscan genomes.

The three insect taxa, *Drosophila yakuba* (Clary and Wolstenholme 1985), *Anopheles gambiae* (Beard et al. 1993) and *Apis mellifera* (Crozier and Crozier 1993), have an identical genome organization with 9 genes encoded on one strand and 4 encoded on the other (figure 4.4). The replication mechanism is known for *Drosophila* (Clary and Wolstenholme 1985) and presumed to be shared by all three taxa. For this log-linear analysis, the three distance classes are defined beginning immediately after the AT rich non-coding region where the first origin of replication is located (figure 4.4). Distance class 1 contains only genes encoded on one strand. There is an extreme bias against both G and C in the arthropod genomes. Because fourfold degenerate sites from *Drosophila* are 94.1% A+T, *Anopheles* are 94.0% A+T and *Apis* are 96.9% A+T, G and C are dropped from all

log-linear analyses of these taxa. The complete *Drosophila* data table has 2 (S) x 3 (D) x 8 (C) x 2 (B) = 96 cells, 1x1x8x2=16 of which are fixed zeros. From the *Anopheles* genome, there is insufficient data for leucine codons and for the *Apis* genome, there are too few arginine codons. These rows of the data table are omitted from their respective log-linear analysis. Consequently, the *Anopheles* and *Apis* data tables are reduced to 2x3x7x2=84 cells, 1x1x7x2=14 of which are fixed zeros. Direct comparison of G^2 values for a single model across insect taxa are inadvisable because of these differences in associated degrees of freedom.

Figure 4.2. Linearized map of the *Cepaea nemoralis* genome (Terrett et al. 1995) illustrating the boundaries of the four classes of the categorical variable (D).

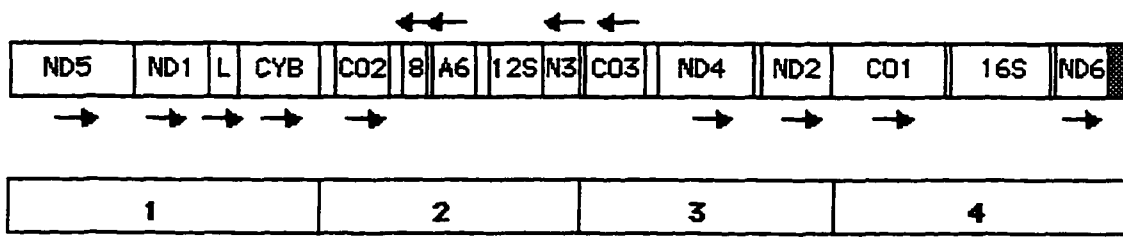


Figure 4.3. Linearized map of the *Katharina tunicata* genome (Boore and Brown 1994) illustrating the boundaries of the four classes of the categorical variable (D).

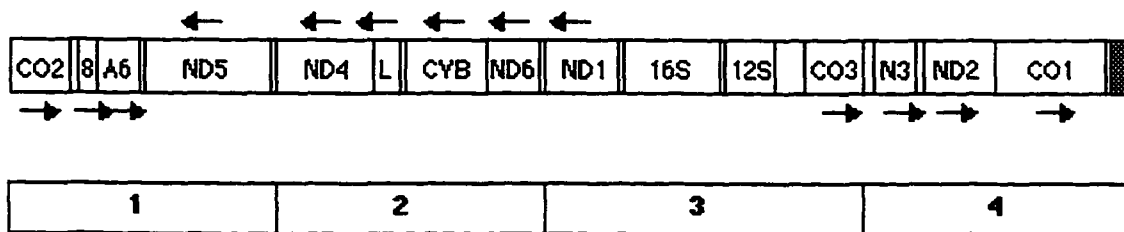
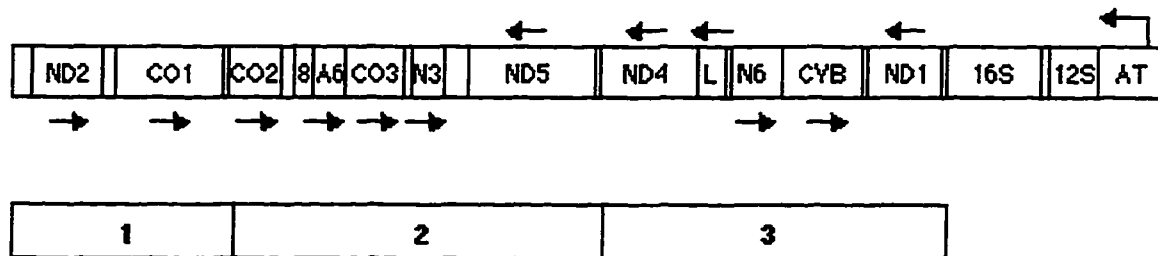


Figure 4.4. Linearized map of the *Drosophila yakuba* genome (Clary and Wolstenholme 1985) illustrating the boundaries of the three classes of the categorical variable (D).



Results and Discussion

Katharina tunicata

The overall goodness-of-fit of each log-linear model to the *Katharina* data (table 4.1) is assessed by the log-likelihood ratio, G^2 , which is asymptotically distributed as X^2 with degrees of freedom equal to the difference between the number of cells in the table and the number of parameters estimated. Table 4.1 also contains two additional criteria for interpreting the fit of each model. The *adjusted-R*² is the amount of the total variation in the data explained by the model. The *AIC* is an index that is minimized for the model that provides the most information about the data and is useful for determining when addition of a term increases the explanatory power of the model.

Under the simplest model considered, [CSD][B], base composition of *Katharina* fourfold degenerate sites can be described as a single multinomial distribution, independent of codon family, strand and position in the genome. This model is rejected based on the G^2 . The extremely high *AIC* is further evidence that this model provides a poor description of the *Katharina* data. Thus, it is unlikely that fourfold degenerate site base composition is

determined by a strand-independent directional mutation pressure with no contextual effects.

Table 4.1. Summary data for log-linear models fit to the *Katharina tunicata* data.

MODEL	G^2	df	p	AIC	$adj.-R^2$
[CSD][CB][SB][DB]	118.52	108	0.2300	142.52	0.8829
[CSD][CB][DB]	261.34	111	0.0000	291.34	0.7322
[CSD][CB][SB]	127.21	117	0.2444	169.21	0.8592
[CSD][SB][DB]	319.70	129	0.0000	385.70	0.5788
[CSD][CB]	385.05	120	0.0000	433.05	0.5561
[CSD][DB]	474.05	132	0.0000	546.05	0.3442
[CSD][SB]	328.68	138	0.0000	412.68	0.4948
[CSD][B]	614.39	141	0.0000	704.39	0.0000

The model including all two-way interaction terms [CSD][CB][SB][DB] ($p=0.2300$) has the lowest AIC of any model tested and explains 88.29% of the total variation in the data. Removal of the distance*base interaction term leads to a slight improvement in fit according the G^2 statistic ($p=0.2444$) and although the AIC value is slightly higher for the [CSD][CB][SB] model than for the [CSD][CB][SB][DB] model, the reduced model explains almost the same level of total variation ($adjusted-R^2=85.29\%$). Removal of either the codon*base or strand*base interaction term results in an unacceptable change in G^2 ($p<0.0001$ for both models). The AIC and $adjusted-R^2$ indicate that the codon*base interaction term carries more information about the data and explains a greater proportion of the total variation than the strand*base interaction term. However, the model that contains only the codon*base two-way interaction term (in addition to fixed terms) is unacceptable based on G^2 , has a high AIC and explains only 55.61% of the total variation.

Table 4.2. The significance of two-way interaction terms involving base for log-linear models of *Katharina tunicata* fourfold degenerate sites.

Model	G^2	df	p
[SCD][CB][DB]	261.34	111	
[SCD][SB][CB][DB]	118.52	108	
[SB]	142.82	3	$p<0.001$
[SCD][CB][SB]	127.21	117	
[SCD][SB][CB][DB]	118.52	108	
[DB]	8.69	9	$0.25<p<0.50$
[SCD][DB][SB]	319.70	129	
[SCD][SB][CB][DB]	118.52	108	
[CB]	201.18	21	$p<0.001$

Table 4.2 shows the significance of each of the two-way interaction terms involving base as evaluated by the change in G^2 when each term is removed from this model. Distance*base is the only two-way interaction term that is not statistically significant at the $p<0.001$ level. The best and simplest model appears to be [CSD][CB][SB] indicating that base composition varies among codon families and between strands in this genome but given strand and codon family, base composition is independent of position in the genome. Table 4.3 shows the *Katharina* composition cross-classified by codon family and strand.

Several evolutionary models are consistent with the intramolecular compositional patterns in *Katharina*. We have already ruled out a simple mutation pressure model. A strand-dependent mutational matrix should not produce variation among codon families within a strand, so it is necessary to invoke higher-order mutational biases to account for the observed patterns with a strictly mutational model. Translational level selection can also produce variation among codon families, but acting alone should lead to the same composition on both strands for a given codon family. Rejection of the [CSD][CB] model,

and conditional independence of base and strand, given codon family, suggests that translational level selection alone is not responsible for *Katharina* base composition. A combination of translational level selection and strand-specific mutation pressure could produce the observed patterns.

Table 4.3. *Katharina tunicata* fourfold degenerate site base composition for each codon family on each strand, followed by the unweighted average of the 8 codon families, and the overall total of each strand.

Codon Family	Strand	%G	%A	%T	%C
ala	1	18.0	16.8	61.0	3.4
	2	2.5	45.1	29.5	23.0
arg	1	33.3	36.4	27.3	3.0
	2	8.3	58.3	8.3	25.0
gly	1	46.5	23.2	21.8	8.4
	2	20.6	41.2	17.5	20.6
leu	1	7.7	25.6	64.1	2.6
	2	1.8	47.6	32.1	18.4
pro	1	4.7	9.4	79.7	6.2
	2	0.0	42.4	40.9	16.7
ser	1	8.9	11.1	76.7	3.3
	2	1.8	38.2	40.9	19.1
thr	1	5.0	25.0	70.0	0.0
	2	4.2	47.9	31.9	16.0
val	1	22.0	21.3	53.3	3.3
	2	1.0	61.0	27.0	11.0
average	1	18.3	21.1	56.7	3.8
	2	5.0	47.7	28.5	18.7
total	1	21.4	19.9	54.3	4.4
	2	4.5	46.8	30.5	18.2

Cepaea nemoralis

Table 4.4 lists the G^2 , *adjusted-R*² and *AIC* for each of the log-linear models fit to the *Cepaea nemoralis* data. The model of complete independence [CSD][B] is acceptable (but barely, $p=0.0549$) based on the G^2 . This model however, has the highest *AIC* value of any model tested. An *adjusted-R*² cannot be calculated since this is the smallest possible model in the analysis and hence serves as a base-line for comparison of all other models. The model including all two-way interaction terms [CSD][CB][SB][DB] ($p=0.3488$) has a low *AIC* and explains 59.32% of the total variation in the data.

Table 4.4. Summary data for log-linear models fit to the *Cepaea nemoralis* data.

MODEL	G^2	df	p	<i>AIC</i>	<i>adj.-R</i> ²
[CSD][CB][SB][DB]	113.13	108	0.3488	137.13	0.5932
[CSD][CB][DB]	120.50	111	0.2532	150.50	0.5507
[CSD][CB][SB]	124.10	117	0.3091	166.10	0.5002
[CSD][SB][DB]	150.79	129	0.0921	216.79	0.2771
[CSD][CB]	131.75	120	0.2184	179.75	0.4473
[CSD][DB]	158.01	132	0.0609	230.01	0.2046
[CSD][SB]	161.87	138	0.0806	245.87	0.0946
[CSD][B]	168.85	141	0.0549	258.85	0.0000

Table 4.5, showing the significance of each of the two-way interaction terms involving base, indicates that the distance*base interaction term is the most expendable, followed by the strand*base interaction. The *AIC* and *adjusted-R*² (table 4.4) also support the idea that the codon*base term is the most important two-way interaction involving base. Furthermore, G^2 , *AIC* and *adjusted-R*² indicate that it is considerably better to remove both the strand*base and distance*base interactions than to remove the codon*base interaction alone. Thus, within a given codon family, there is strong support for the independence of

base composition and both position in the circular genome and strand. It is not inconceivable that a single multinomial distribution could describe the base composition of all *Cepaea* fourfold degenerate sites; However, a model that allows composition to vary among codon families fits the data much better.

Table 4.5. The significance of two-way interaction terms involving base in log-linear models of *Cepaea nemoralis* fourfold degenerate sites.

Model	G^2	df	p
[SCD][CB][DB]	120	111	
[SCD][SB][CB][DB]	113.13	108	
[SB]	6.87	3	$0.05 < p < 0.10$
[SCD][CB][SB]	124.1	117	
[SCD][SB][CB][DB]	113.13	108	
[DB]	10.97	9	$0.25 < p < 0.50$
[SCD][DB][SB]	150.79	129	
[SCD][SB][CB][DB]	113.13	108	
[CB]	37.66	21	$p < 0.025$

In terms of evolutionary models, this means that *Cepaea* base composition may be determined entirely by a strand-independent mutational pressure with no contextual biases, but if the biases are the same for both strands, the relative frequency of A should equal T and G should equal C (Sueoka 1995). The fourfold degenerate site composition of the *Cepaea* genome (Table 4.6) shows that overall G and C are nearly equal, but that T is greater than A, for the total data, the weighted average of codon families, and all but one individual codon families. This bias seems inconsistent with a no strand-bias mutation only model. Allowing for simple strand-specific mutation pressures does not explain the increase in explanatory power gained by allowing composition to vary among codon

families. Translational level selection alone could explain the variation among codon families and would not necessarily lead to rejection of the independence of base and strand if the distribution of amino acids is similar for both strands. However, correlations of fourfold degenerate site composition and the composition in other regions of the genome are indicative of directional mutation pressures (Chapter V) so it is unlikely that selection, if it occurs, is acting in an unbiased mutational background. Higher-order mutational biases could also produce the variation among codon families.

Table 4.6. *Cepaea nemoralis* fourfold degenerate site base composition for each codon family, followed by the unweighted average of the 8 codon families, and the overall total.

Codon Family	%G	%A	%T	%C
ala	16.15	26.54	36.15	21.15
arg	23.66	20.43	32.26	23.66
gly	20.56	18.69	34.11	26.64
leu	19.50	31.89	31.58	17.03
pro	17.12	22.60	38.36	21.92
ser	17.68	20.73	36.59	25.00
thr	17.22	31.67	36.11	15.00
val	19.93	29.54	35.94	14.59
average	18.98	25.26	35.14	20.62
total	18.78	26.37	34.98	19.87

Mollusc summary

Katharina tunicata and *Cepaea nemoralis* fourfold degenerate site base compositions share some characteristics. Both are greater than 50% AT. Overall, the frequency of T is greater than A. Both T and A are greater than G or C, which are roughly equal. However, intramolecular compositional patterns and the mechanisms generating these patterns, have

clearly diverged since these two taxa last shared a common ancestor over 500 MYA (Terrett et al. 1995). *Katharina* exhibits compositional variation between strands, while *Cepaea* composition is not strand-dependent. This may indicate that the *Katharina* mutational matrix is asymmetric, but the *Cepaea* matrix is symmetric. However, variation among codon families shows that a strictly mutational model does not adequately explain the compositional pattern of either genome unless the mutation pattern for a given nucleotide site is context dependent. Translational level selection is a viable alternate explanation for the variation among codon families observed in both molluscs.

Notably, for both molluscan genomes the base composition is independent of the position in the molecule. The mammalian genomes analyzed in Chapter III showed a strong relationship between base composition and discrete distance classes. Predicted values for the best log-linear model supported the idea that there is a compositional gradient around mammalian genomes compatible with predictions of a mutational matrix that varies with time spent single-stranded during replication. There is no evidence of a similar gradient in the molluscan genomes. This could indicate any of several things:

1. There is no real compositional gradient in mammals or molluscs. Mutation may be strand specific for all these taxa, except *Cepaea*, but not because one strand remains unprotected during replication while the other is always paired. The strong distance*base relationship in mammals arises for some other unknown reason.
2. The compositional gradient is real in mammals and does reflect the difference between mutational patterns for double and single-stranded DNA. The independence of base and distance class in molluscs indicates that either the replication mechanism does not involve extended periods in a single-stranded state or the exposed strand is protected. In the absence of direct experimental data for

either mollusc, it is impossible to determine whether there is a fundamental difference in the pattern or components of replication.

3. This analysis is unable to detect compositional gradients in the mollusc genomes, despite a position dependent mutational pattern. There have been numerous genome rearrangements since these taxa diverged. In contrast, the mammalian genome organization is identical to the lamprey gene order, suggesting that this architecture has been stable for nearly 500 million years. Periodical changes in gene order would shuffle the base composition among distance classes. If the composition has not had sufficient time to attain a new mutational equilibrium, the gradient would be obscured.

Drosophila yakuba

Table 4.7 lists the G^2 , df and associated p -value for each log-linear model fit to the *Drosophila yakuba* data. With the elimination of G and C from the *Drosophila* data table, as described in the methods section, it is possible to consider the fit of models containing three-way interaction terms. This increases the total number of models to 19 and in tabular form it is difficult to see the relationship between models that provide an acceptable fit and those that are rejected. Figure 4.5 is a path diagram illustrating the relationships between the 19 possible models.

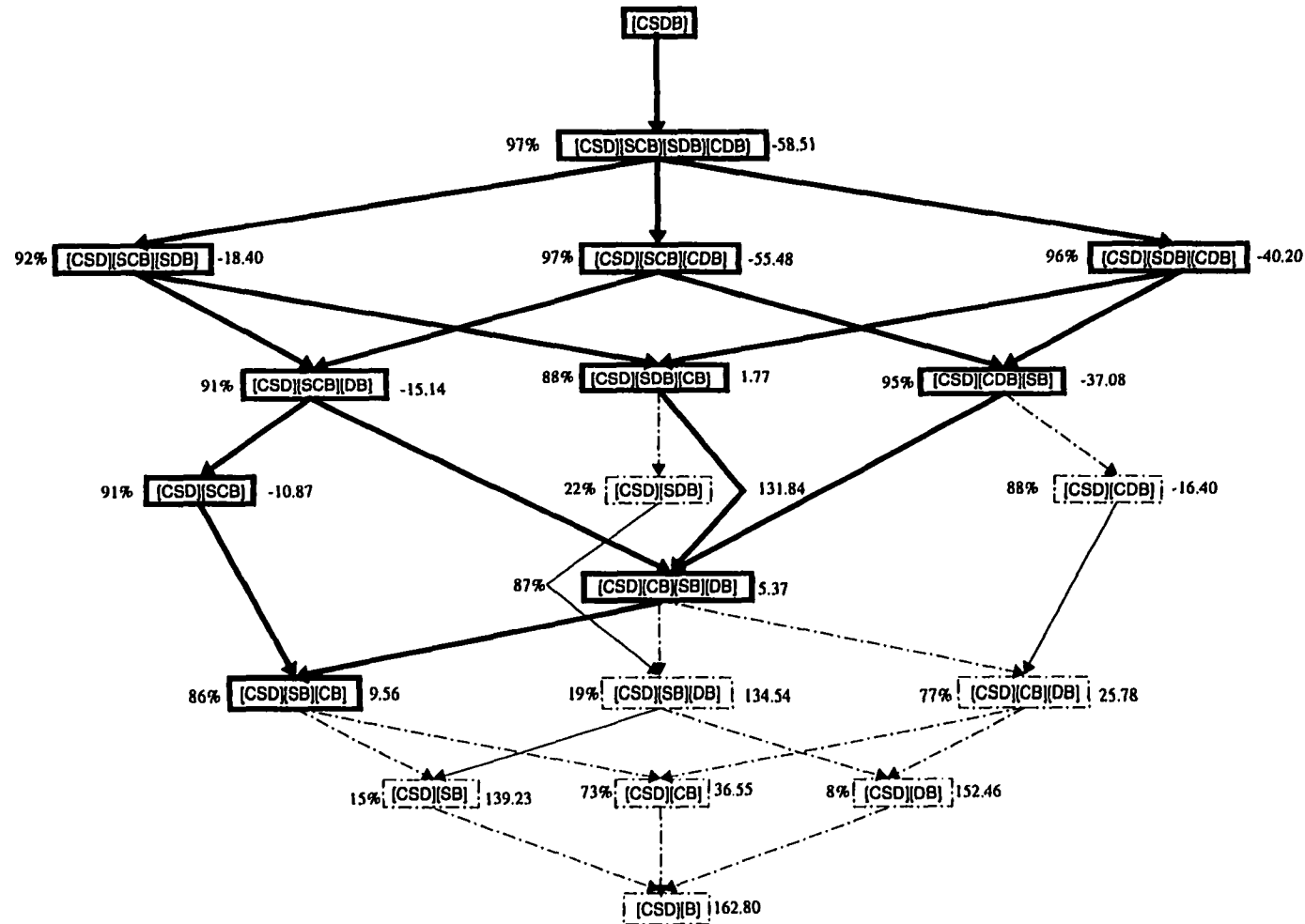
If we begin with the fully saturated model [CSDB] at the top of the path diagram and move through the models step-wise, removing one term at a time until the overall goodness-of-fit is rejected, we converge at the model [CSD][SB][CB]. Base composition varies among codon families and between strands in this genome but given strand and codon family, the base composition is independent of position in the genome. This model explains 86% of the variation in the *Drosophila* data table and has a low AIC . The change

in G^2 upon removal of either the codon*base or strand*base interaction term shows that both are significant at the $p=0.001$ level. The AIC and *adjusted-R*² indicate that the codon*base interaction term is far more informative than the strand*base interaction term. Conditional independence of base and strand [CSD][CDB], or of base and codon family [CSD][SDB], are both rejected for the *Drosophila* data even when the three way interaction term including the remaining two variables and base is added to the models. Addition of the three-way interaction term strand*codon*base to the [CSD][SB][CB] model increases the explanatory power very little.

Table 4.7. Summary data for log-linear models fit to the *Drosophila yakuba* data.

MODEL	G^2	df	p
[CSD][CSB][SDB][CDB]	7.49	7	0.3795
[CSD][SDB][CDB]	11.80	14	0.6221
[CSD][CSB][SDB]	19.60	21	0.5464
[CSD][CSB][CDB]	8.52	8	0.3842
[CSD][SDB][CB]	25.77	28	0.5854
[CSD][CSB][DB]	20.86	22	0.5293
[CSD][CDB][SB]	12.92	15	0.6089
[CSD][SDB]	141.84	35	0.0000
[CSD][CSB]	21.13	24	0.6312
[CSD][CDB]	31.60	16	0.0113
[CSD][CB][SB][DB]	27.37	29	0.5517
[CSD][CB][DB]	45.78	30	0.0326
[CSD][CB][SB]	27.56	31	0.6438
[CSD][SB][DB]	142.54	36	0.0000
[CSD][CB]	52.55	32	0.0125
[CSD][DB]	158.46	37	0.0000
[CSD][SB]	143.23	38	0.0000
[CSD][B]	164.80	39	0.0000

Figure 4.5. Path diagram illustrating the relationships between 19 log-linear models fit to the *Drosophila yakuba* data. Each boxed model is connected by a line to every other model that differs by inclusion of a single term. Models in bold boxes provide an adequate fit to the *Drosophila* data. Overall goodness-of-fit is rejected at the 0.05 level for models shown in dashed boxes. The significance of individual terms is evaluated by examining the change in G^2 upon deletion of that term from a model. Solid lines connecting models indicate that the removed term has an associated p -value > 0.05 and dashed lines indicate that the removed term has an associated p -value < 0.05 . The percentage shown to the left of each model is the *adjusted-R*² and the number to the right of each model is the *AIC*.



Overall, the intramolecular compositional patterns in *Drosophila* are reminiscent of those described earlier for *Katharina*. A simple mutation pressure should result in complete independence of base from the remaining variables. We reject complete independence. Either a strand-specific mutational bias model including contextual effects or a combination of mutational bias and translational level selection are consistent with the simplest adequate log-linear model.

The lack of a significant distance*base interaction is similar to both molluscs and again, unlike the position dependent compositional effect in mammals. *Drosophila* is one of the few non-mammalian taxa where there is direct experimental evidence for a replication mechanism. As in mammals, replication begins at a site in the major non-coding region, and continues unidirectionally copying one strand. In mammals, the replication of the second strand begins after the first strand is extended two-thirds of the way around the genome, exposing a second origin of replication. In *Drosophila*, the second origin is located in the same major non-coding region as the first origin, so replication of one strand is nearly complete before replication of the other is initiated. Thus, mitochondrial DNA replication in *Drosophila* is even more asymmetric than in mammals, leaving regions of one strand in a single stranded state for an entire replication cycle. Many researchers have suggested that the compositional variation between strands is related to differences in mutational spectra for single vs. double-stranded DNA and that this mutational pressure should result in compositional gradients. We observed that a simple log-linear model for mammalian fourfold degenerate site composition predicted gradients consistent with this hypothesis. The complete lack of significance of the distance*base interaction term in log-linear models for *Drosophila* can not be attributed to an unknown and possibly symmetric replication mechanism.

It is surprising to find that the base composition is strand-specific in light of our earlier observation that fourfold degenerate sites from *Drosophila yakuba* exhibited the least skew

of any mitochondrial genome analyzed (Perna and Kocher 1995b, Chapter II). The total composition of each strand shown in table 4.8 reveals the reason for this apparent inconsistency. The earlier analysis considered genes from strand 1 only and tested for skew by examining the difference between the frequency of *A* and *T* (*G* and *C*) on that strand. For strand 1, these frequencies are nearly equal, hence we detected no skew. This log-linear analysis examines a different set of sites from each strand. Had the strand 2 sites been used in the earlier analysis, we would have concluded that *Drosophila yakuba* exhibits a pattern of skew similar to that observed in vertebrates and *Apis*, the only other insect considered in that analysis. The simplest evolutionary model that accounts for observed intramolecular compositional patterns requires either contextual effects or translation level selection in addition to simple mutational biases. Under these conditions, the samples from each strand no longer represent estimates of either the same distribution or one distribution and its mirror image, both of which are expected to result in consistent estimates of skew for the two strands. Clearly the complexity of intramolecular compositional patterns serves as a warning about the wisdom of drawing very specific conclusions about patterns of mitochondrial evolution from simple estimates of composition.

Table 4.8. *Drosophila yakuba* fourfold degenerate site base composition for each codon family on each strand, followed by the unweighted average of the 8 codon families, and the overall total of each strand.

Codon Family	Strand	%G	%A	%T	%C
ala	1	0.9	23.7	69.3	6.1
	2	1.7	17.0	78.0	3.4
arg	1	0.0	89.5	10.5	0.0
	2	28.6	52.4	19.0	0.0
gly	1	5.9	70.4	23.7	0.0
	2	16.5	40.0	41.2	2.4
leu	1	0.0	31.7	63.4	4.9
	2	11.1	33.3	55.6	0.0
pro	1	2.1	40.2	54.6	3.1
	2	3.0	18.2	78.8	0.0
ser	1	2.0	49.0	47.0	2.0
	2	0.0	36.1	62.6	1.2
thr	1	0.7	49.0	48.2	2.1
	2	2.3	34.1	63.6	0.0
val	1	2.5	52.5	44.2	0.8
	2	6.8	40.5	50.0	2.7
average	1	1.8	50.8	45.1	2.4
	2	8.8	34.0	56.1	1.2
total	1	2.1	49.5	46.1	2.3
	2	7.2	34.0	57.1	1.7

Anopheles gambiae

Table 4.9 lists the G^2 , df and associated p -value for each log-linear model fit to the *Anopheles gambiae* data. Figure 4.6 is a path diagram for the 19 possible *Anopheles* models analogous to figure 4.5 for *Drosophila*. The simplest model that adequately fit the *Anopheles* data is [CSD][SCB]. Although this model differs from the simplest adequate model for the *Drosophila* data by the inclusion of a three-way interaction term, the conditional independence interpretation is the same. Base composition varies among codon families and between strands in this genome but given strand and codon family, the base composition is independent of position in the genome. This model explains 87.45% of the variation in the *Anopheles* data table and has a low AIC . The strand*codon*base interaction term itself, is not significant at the 0.05 level, but removal of this term from the [CSD][SCB] model leads to an unacceptable overall fit. Careful inspection of the path diagram reveals that the codon*base and strand*base terms are significant at the 0.05 level (models connected by dashed lines) regardless of whether or not third order interaction terms are included. The codon*base interaction term always increases $adjusted-R^2$ and decreases the AIC more than the strand*base interaction term, but their relative informativeness is much more equal for the *Anopheles* data than for the *Drosophila* data. Conditional independence of base and strand or of base and codon family, are both rejected for the *Anopheles* data even when the three way interaction term including the remaining two variables and base is added to the models.

Table 4.10 shows the relative frequency of the four bases at *Anopheles* fourfold degenerate sites cross-classified by codon family and strand. Note that the near mirror image relationship between total composition for the two strands suggests that a simple strand-specific mutational pressure could describe *Anopheles* data, but the additional variation among codon families indicates that this model is inadequate. Again, we conclude that the simplest evolutionary model that provides for the observed intramolecular variation

includes either dinucleotide mutational biases or translational level discrimination among synonymous codons.

Table 4.9. Summary data for log-linear models fit to the *Anopheles gambiae* data.

MODEL	G^2	df	p
[CSD][CSB][SDB][CDB]	3.06	6	0.8018
[CSD][SDB][CDB]	12.21	12	0.4290
[CSD][CSB][SDB]	24.76	18	0.1317
[CSD][CSB][CDB]	4.47	7	0.7240
[CSD][SDB][CB]	35.94	24	0.0556
[CSD][CSB][DB]	28.05	19	0.0825
[CSD][CDB][SB]	13.41	13	0.4165
[CSD][SDB]	110.94	30	0.0000
[CSD][CSB]	29.50	21	0.1024
[CSD][CDB]	59.14	14	0.0000
[CSD][CB][SB][DB]	39.24	25	0.0349
[CSD][CB][DB]	83.81	26	0.0000
[CSD][CB][SB]	40.64	27	0.0446
[CSD][SB][DB]	114.05	31	0.0000
[CSD][CB]	103.18	28	0.0000
[CSD][DB]	154.50	32	0.0000
[CSD][SB]	115.81	33	0.0000
[CSD][B]	172.68	34	0.0000

Figure 4.6. Path diagram illustrating the relationships between 19 log-linear models fit to the *Anopheles gambiae* data. Each boxed model is connected by a line to every other model that differs by inclusion of a single term. Models in bold boxes provide an adequate fit to the *Anopheles* data. Overall goodness-of-fit is rejected at the 0.05 level for models shown in dashed boxes. The significance of individual terms is evaluated by examining the change in G^2 upon deletion of that term from a model. Solid lines connecting models indicate that the removed term has an associated p -value > 0.05 and dashed lines indicate that the removed term has an associated p -value < 0.05 . The percentage shown to the left of each model is the *adjusted-R*² and the number to the right of each model is the AIC.

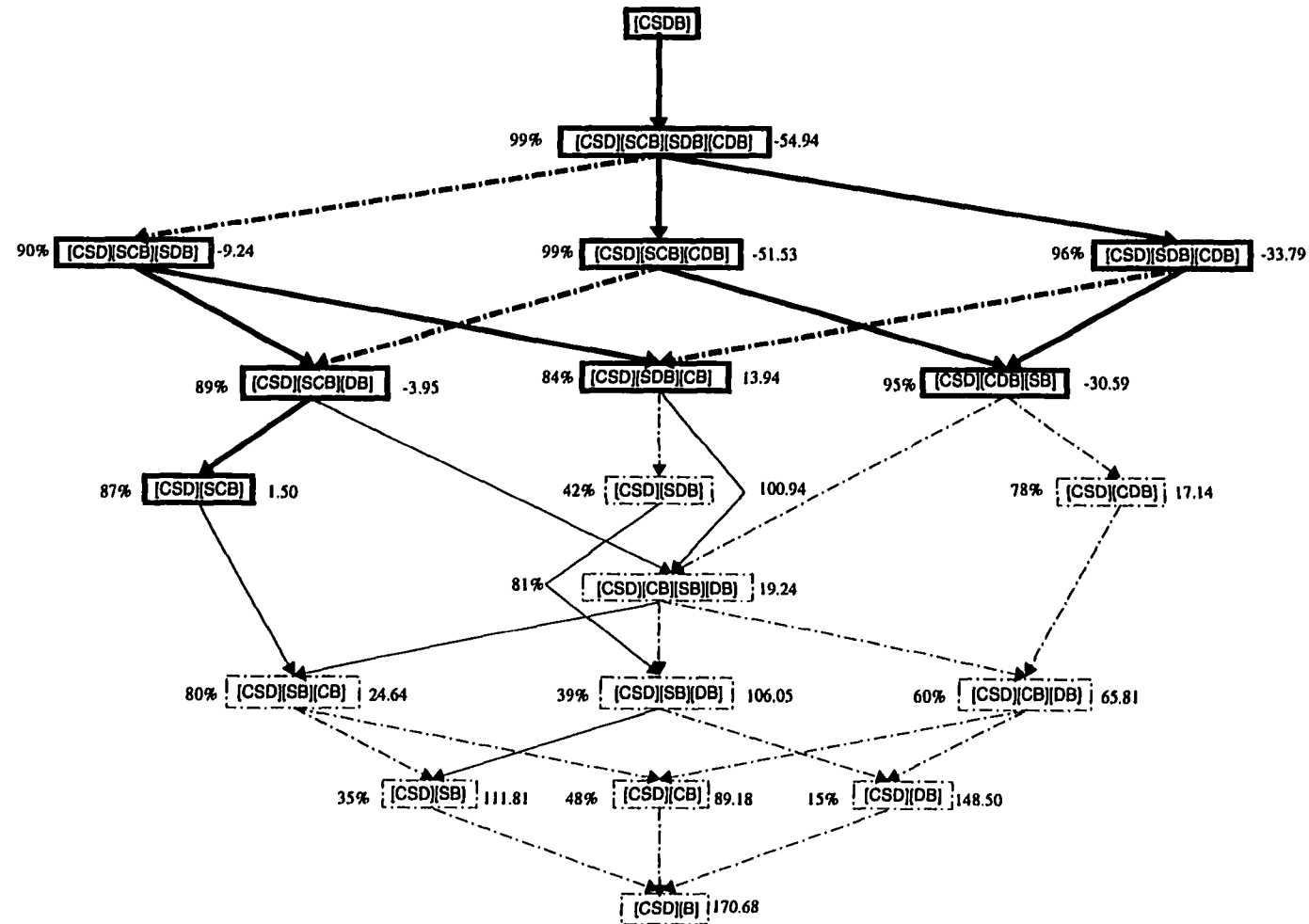


Table 4.10. *Anopheles gambiae* fourfold degenerate site base composition for each codon family on each strand, followed by the unweighted average of the 8 codon families, and the overall total of each strand. Asterisks indicate where small sample size is likely to lead to large errors.

Codon Family	Strand	%G	%A	%T	%C
ala	1	0.0	37.4	53.9	8.7
	2	0.0	26.2	70.5	3.3
arg	1	5.3	92.1	2.6	0.0
	2	15.0	60.0	25.0	0.0
gly	1	4.5	78.2	15.8	1.5
	2	22.4	38.8	35.3	3.5
leu	1	1.8	58.9	37.5	1.8
	2	0.0	46.7	53.3	0.0
pro	1	0.0	41.6	55.4	3.0
	2	3.0	30.3	66.7	0.0
ser	1	3.2	53.6	40.8	2.4
	2	1.2	16.9	79.5	2.4
thr	1	0.0	50.6	47.4	1.9
	2	4.1	26.5	67.4	2.0
val	1	2.5	60.2	36.4	0.8
	2	2.3	39.5	55.8	2.3
average	1	2.2	59.1	36.2	2.5
	2	4.7	32.1	61.2	1.9
total	1	1.9	56.3	39.1	2.7
	2	6.5	32.2	59.0	2.3

As in *Drosophila*, log-linear models including terms in addition to those absolutely necessary to provide an adequate overall fit, increase the *adjusted-R*² and decrease the *AIC*. Further inspection of the path diagram reveals a feature unique to the *Anopheles* log-linear analysis. The codon*distance*base interaction is significant ($0.025 > p > 0.01$) when it is

evaluated by the difference in G^2 upon removal from each of the five models including the term. Consider whether the translational level selection model or the dinucleotide mutation model is more likely to lead to a significant codon*distance*base interaction. The level of expression is not likely to vary considerably among mitochondrial genes, so the strength of selection on fourfold degenerate sites should be comparable for any subset of genes. The variation in the data attributed to a codon*distance*base interaction could be explained by differences in the proportion of a codon family on each strand between distance classes. However, the significance of this term remains constant even when we include strand*base, strand*codon*base, or strand*distance*base terms. Under a contextual bias mutational model, a significant codon*distance*base interaction might arise as a result of variation among distance classes in the composition of positions adjacent to particular codon families.

Apis mellifera

Table 4.11 lists the G^2 , df and associated p -value for each log-linear model fit to the *Apis mellifera* data. Figure 4.7 is a path diagram for the 19 possible *Apis* models. Comparison with the analogous path diagram from *Drosophila* (figure 4.5) reveals that the same log-linear models that provided an adequate overall fit to the *Drosophila* data are acceptable for the *Apis* data. Again, the simplest model that adequately fit is the conditional independence of distance class and base model, [CSD][SB][CB]. This model explains 74.67% of the variation in the *Apis* data table and has a low AIC . This R^2 is roughly 10% less than the *adjusted-R²* for the simplest models that adequately fit either *Drosophila* or *Anopheles*, and to achieve a comparable level of explanatory power, at least one higher order interaction term must be included in the model. The change in G^2 upon removal of either the codon*base or strand*base interaction term shows that both are significant at the

$p=0.001$ level. As in *Anopheles*, the codon*base interaction term carries only slightly more explanatory power than the strand*base interaction term. All models consistent with conditional independence of base and strand or of base and codon family are rejected.

Table 4.11. Summary data for log-linear models fit to the *Apis mellifera* data.

MODEL	G^2	df	p
[CSD][CSB][SDB][CDB]	6.07	6	0.4155
[CSD][SDB][CDB]	13.45	12	0.3371
[CSD][CSB][SDB]	18.76	18	0.407
[CSD][CSB][CDB]	6.17	7	0.5197
[CSD][SDB][CB]	28.01	24	0.2594
[CSD][CSB][DB]	19.89	19	0.4013
[CSD][CDB][SB]	13.84	13	0.3852
[CSD][SDB]	63.13	30	0.0004
[CSD][CSB]	21.82	21	0.4098
[CSD][CDB]	44.23	14	0.0001
[CSD][CB][SB][DB]	29.54	25	0.2420
[CSD][CB][DB]	59.22	26	0.0002
[CSD][CB][SB]	31.26	27	0.2604
[CSD][SB][DB]	64.57	31	0.0004
[CSD][CB]	64.89	28	0.0001
[CSD][DB]	97.35	32	0.0000
[CSD][SB]	66.48	33	0.0005
[CSD][B]	103.31	34	0.0000

Figure 4.6. Path diagram illustrating the relationships between 19 log-linear models fit to the *Apis mellifera* data. Each boxed model is connected by a line to every other model that differs by inclusion of a single term. Models in bold boxes provide an adequate fit to the *Apis* data. Overall goodness-of-fit is rejected at the 0.05 level for models shown in dashed boxes. The significance of individual terms is evaluated by examining the change in G^2 upon deletion of that term from a model. Solid lines connecting models indicate that the removed term has an associated p -value > 0.05 and dashed lines indicate that the removed term has an associated p -value < 0.05 . The percentage shown to the left of each model is the *adjusted-R*² and the number to the right of each model is the *AIC*.

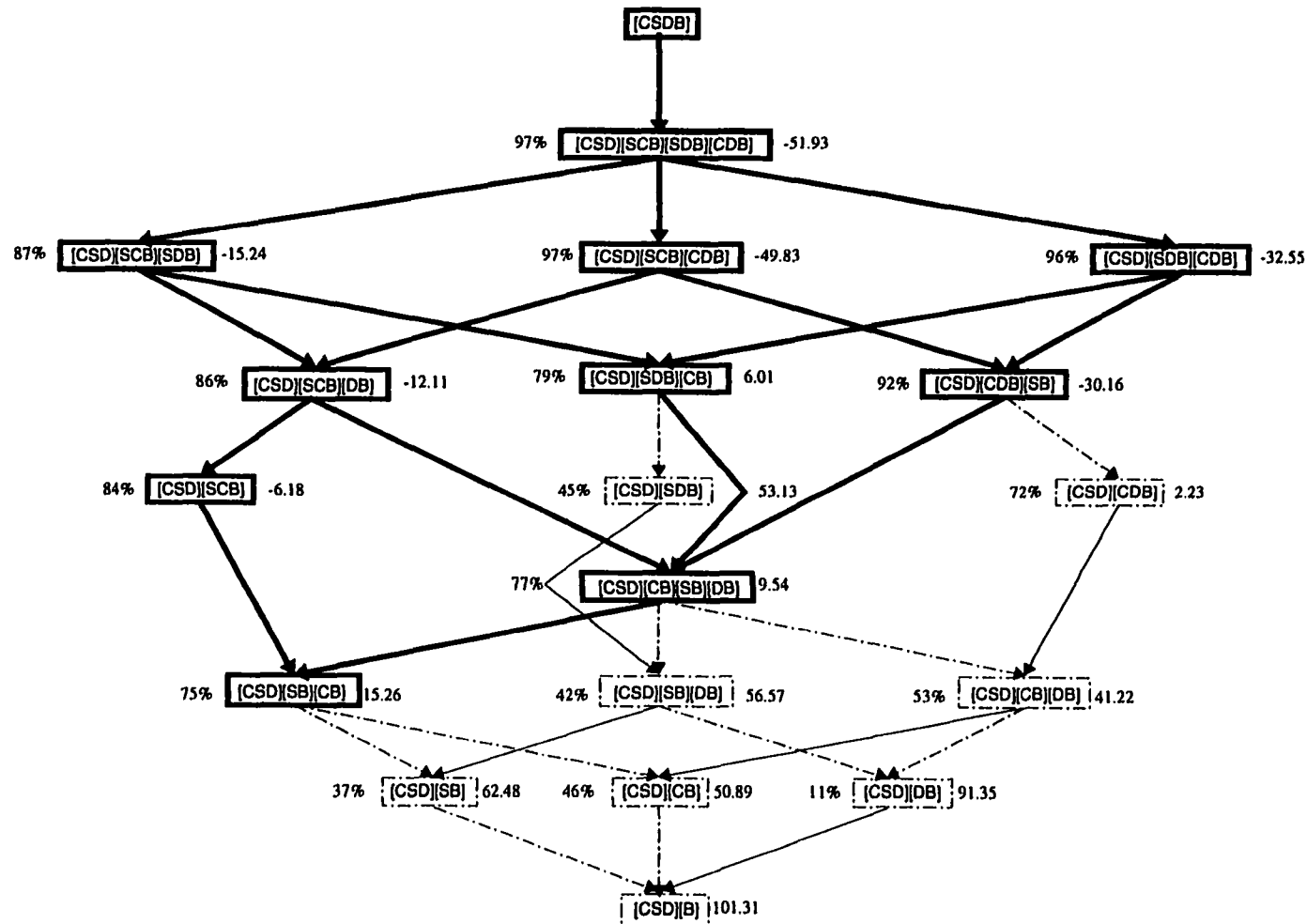


Table 4.12 shows the relative frequency of the four bases at *Apis* fourfold degenerate sites cross-classified by codon family and strand. As in *Drosophila*, the total base composition of the samples from each strand are neither identical nor mirror images. Strand 1 data from *Apis* were included in our previous analysis of skew in metazoan mitochondrial genomes, and the difference between the frequency of *A* and *T* indicated a strong strand-specific compositional distribution. Had we used the strand 2 sites instead, we would have concluded that *Apis* exhibited very little *AT-skew*. Although the situation is similar to that observed for the *Drosophila*, it is important to note that while strand 1 showed no skew and strand 2 exhibited skew in that genome, here strand 1 is skewed but strand 2 is not. Again, we attribute this difference between samples from the two strands to an evolutionary model that involves either contextual effects in the mutational matrix or translation level natural selection.

Table 4.12. *Apis mellifera* fourfold degenerate site base composition for each codon family on each strand, followed by the unweighted average of the 8 codon families, and the overall total of each strand. Asterisks indicate where small sample size is likely to lead to large errors.

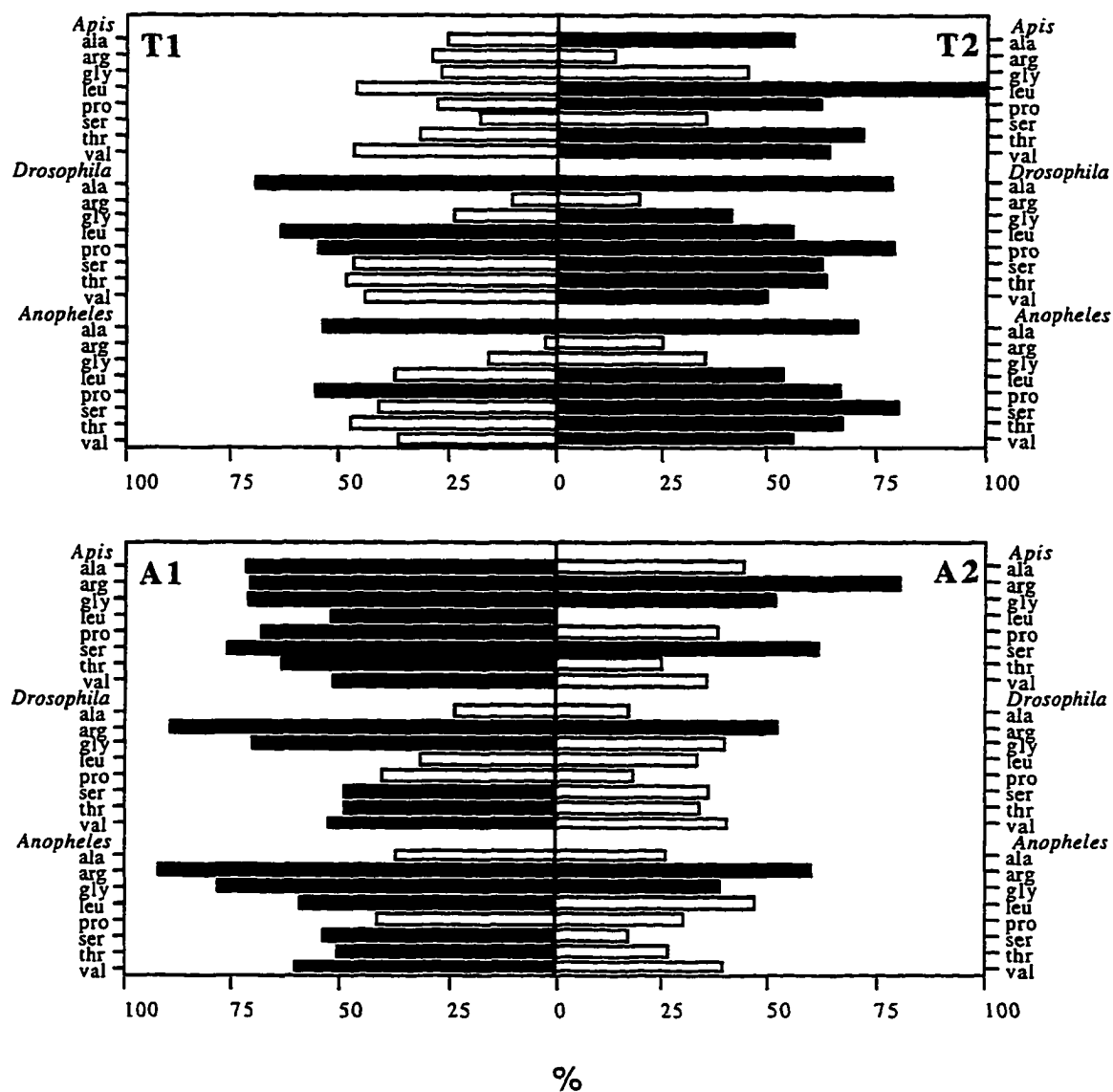
Codon Family	Strand	%G	%A	%T	%C
ala	1	0.0	71.8	25.6	2.6
	2	0.0	44.4	55.6	0.0
arg	1	0.0	70.8	29.2	0.0
	2	6.7	80.0	13.3	0.0
gly	1	1.3	71.4	27.3	0.0
	2	3.4	51.7	44.8	0.0
leu	1	0.0	52.2	46.4	1.4
	2	0.0	0.0	100.0	0.0
pro	1	0.0	68.3	28.0	3.7
	2	0.0	38.1	61.9	0.0
ser	1	0.0	76.1	17.8	6.1
	2	1.5	61.8	35.3	1.5
thr	1	0.0	63.1	34.0	4.8
	2	3.6	25.0	71.4	0.0
val	1	0.0	51.6	44.2	0.8
	2	6.8	40.5	46.8	1.6
average	1	0.2	65.7	31.6	2.4
	2	3.1	48.8	47.0	0.4
total	1	0.2	66.7	29.7	3.4
	2	1.9	47.4	50.4	0.4

Insect Summary

The log-linear analyses of fourfold degenerate sites from all three insects, *Drosophila yakuba*, *Anopheles gambiae* and *Apis mellifera* revealed complex intramolecular compositional patterns. Each analysis individually supports the idea that insect mitochondrial base composition varies between strands and among codon families implicating either contextual effects or natural selection in addition to basic mutational biases. The observed intramolecular variation can obscure efforts to characterize variation among mitochondrial genomes using simple descriptors of base composition.

The two Dipterans, *Drosophila* and *Anopheles*, last shared a common ancestor with the Hymenopteran, *Apis*, about 280 MYA (Crozier and Crozier 1993). The high frequency of *AT* base pairs in insect mitochondrial genomes is widely appreciated, and the fourfold degenerate sites of these three taxa all exceed 93.5% *A+T*. Even among insect taxa, there is variation in *AT* content, and *Apis* is known as the most extreme of all metazoans characterized to date (Crozier and Crozier 1993, Jermin et al. 1994). Our previous analysis of skew in mitochondrial genomes suggested that the compositional variation between *Apis* and *Drosophila* was more complex than a simple increase in net %*AT* and involved a fundamental change in how the *AT* base pairs were oriented on the two strands of these genomes (Perna and Kocher 1995b). This analysis reveals that even measures that consider strand-specific compositional patterns may not tell the whole story. However, it is not a trivial task to compare the composition of genomes using tables cross-classified by codon family and strand (tables 4.8, 4.10 and 4.12). Figure 4.8 is a more visual representation of *T* (upper panel) and *A* (lower panel) usage for the three taxa. Bars on the left side of each panel show the relative frequency of the base on strand 1. Bars on the right side show the relative frequency of each base on strand 2. Shaded bars indicate

Figure 4.8. Comparison of patterns of synonymous codon usage for the three insects. The percentage of synonymous codons ending in *T* on strand 1 (T1), *T* on strand 2 (T2), *A* on strand 1 (A1) and *A* on strand 2 (A2) are shown for each of the eight fourfold degenerate codons from each taxon.



the most frequently used base for the third position of each codon family on each strand.

There are several notable features of figure 4.8:

1. Under a simple *AT* vs. *GC* mutation pressure model, composition of the two strands is identical and the frequency of *T* equals the frequency of *A* within a strand. Each pair of bars should be symmetrical about zero, the left sides of both panels should be identical, and the right sides should also be identical. In short, all bars for a given taxon should be the same length. The log-linear analyses indicated that this evolutionary model is unlikely to account for intramolecular compositional patterns in any of the insects individually. Nor does this model appear to explain the variation among taxa seen in figure 4.8. There is not a universal increase or decrease in either base across all codon families on both strands in *Apis* relative to *Drosophila* and *Anopheles*. Other authors have discussed the compositional variation among these taxa in terms of a symmetrical *AT* mutation pressure (Crozier and Crozier 1993, Jermin and Crozier 1994). While it is indisputable that *Apis* has a higher overall *AT* content than the other two taxa, it is now equally clear that this increase in *AT* content is not randomly distributed among synonymous codon positions.

2. The observation that the most frequently used third position base for any given codon family often varies between strands has important implications for some possible approaches to detecting translation level selection. In either the upper or lower panel of figure 4.8, any codon family shown with a black bar on one side of the graph and a white bar on the other, exhibits this phenomenon. In *Apis*, 5 out of the 8 fourfold degenerate codon families (ala, leu, pro, thr and val) end most frequently in a different base on each of the two strands. Four codon families

exhibit this variation in *Drosophila* (gly, ser, thr, val) and *Anopheles* (leu, ser, thr, val). One means of detecting the effects of translation level selection is to compare ratios of polymorphism within a population and divergence between species for preferred and unpreferred classes of substitution, where preferred indicates that the change results in replacement of a suboptimally translated codon with an optimally translated codon (Akashi, 1995). This approach requires that optimal codons can be identified. At least one group (Dave Rand, personal communication) has applied this method to insect mitochondrial genomes by assigning the most frequent codon optimal status. Figure 4.8 illustrates that this method would assign optimal status to different codons depending on the gene chosen for the analysis. Since there is only one tRNA translating each mitochondrial codon family, and all mitochondrial genes are expressed at high levels, there is no reason to expect that one codon is optimal for one gene, while another codon is better for another gene. Incorrect assignment of optimal codons is a serious limitation for the utility of this approach for mitochondrial DNA. Optimality of mitochondrial codons could be determined experimentally, but to date, very few studies have addressed biochemical parameters of mitochondrial codon-anticodon interactions and none have examined the effect of these interactions on translation using mitochondrial machinery. Abundant evidence from eukaryotic nuclear and prokaryotic genomes supports the idea that natural Watson-Crick base-pairing between the third position of the codon and the corresponding position of the anticodon leads to the greatest binding affinity. However, it is not clear that maximizing the strength of this interaction leads to optimal translation. This, coupled with recent observations of post-transcriptional editing of mitochondrial tRNA's (Yokobori and Paabo 1995) may mean that optimal status can not be assigned to the codon that matches the anticodon of the tRNA gene either.

3. Across all three taxa, there is a consistent difference between strands that is detectable through the noise of variation among codon families. The frequency of *T* on strand 1 is consistently lower than *T* on strand 2 and conversely *A* on strand 1 is higher than *A* on strand 2. Recall that in this analysis, measurements for the two strands are from separate (independent) data, but that mitochondrial DNA is a double-stranded molecule and the *T1* sample can be viewed as a second *A2* sample. In this light, the pattern in figure 4.8 appears to be a conserved overall strand bias across these three insect genomes. One consistent explanation is that for each genome there is an underlying pattern of mutation that differs between strands, perhaps dominated by mutational pathways that target single nucleotides. Variation among codon families results from deviations from this basal pattern due to contextually sensitive mutational pathways or natural selection. The conserved strand-specific pattern would indicate that the underlying mutational spectra are similar for the three taxa.

4. Intramolecular compositional patterns are more similar for *Drosophila* and *Anopheles* than for comparisons of either Dipteran with *Apis*. This is easily seen by comparing the general shape of the set of bars for each taxon in either panel of figure 4.8. This situation might arise because more fourfold degenerate sites remain unmutated between the two dipterans since their more recent divergence. However, these taxa are likely to be well into saturation range for synonymous sites (Crozier and Crozier 1993). If variation among codon families within a genome arises from contextual biases in the mutational spectrum, taxa where the contexts are conserved should show similar patterns. For any given codon family, the dinucleotide on one side of the third codon position will always be same for all taxa

by definition. However, the distribution of dinucleotides on the other side will be determined by the amino acid sequence of the encoded protein. *Drosophila* and *Anopheles* mitochondrial proteins are less divergent and thus, contextual effects are less likely to lead to species-specific synonymous codon usage patterns. The amino acid identity between *Apis* and *Drosophila*, however, ranges from 70% for COI to just 27% for ND2 (Crozier and Crozier 1993). Even if the mutational spectra were identical, the contextual differences between these taxa could change the pattern of variation among codon families. Alternatively, the conserved intramolecular distribution for *Drosophila* and *Anopheles* may reflect a lack of divergence in the mutational patterns or relative fitness of particular codons between these taxa.

Finally, all three insect log-linear analyses support the idea that fourfold degenerate site base composition is independent of position in the molecule. The replication mechanism is known to be asymmetric for *Drosophila*, and is presumed identical in the other two taxa. This mechanism leaves one strand in a single-stranded state for a portion of the replication cycle. Regions of this strand closest to the first origin of replication will be single-stranded longer than regions that are closer to the second origin. This analysis does not reveal any evidence that differences in mutational spectra for single and double-stranded DNA have created a compositional gradient around the insect mitochondrial genomes. Unlike the frequent rearrangements observed between the two molluscs discussed earlier, conservation of gene order among arthropods suggests that there has been sufficient time to achieve mutational equilibrium. If there is a position dependent mutational bias affecting *A* and *T* in insects, it must be subtle enough that compositional differences between the three defined distance classes are not statistically significant.

Conclusions

Fourfold degenerate site base composition varies among codon families in all five invertebrates considered here and the four mammals previously analyzed. Composition also varies between strands in genomes of all taxa except *Cepaea nemoralis*. The simplest evolutionary model consistent with these intramolecular compositional patterns involves either contextual effects or translational level natural selection in addition to simple mutational biases. The complexity of intramolecular compositional patterns described in this analysis suggest that a single stationary substitution matrix may provide a poor representation of the true process acting on even a single lineage.

CHAPTER V

STRAND-SPECIFIC DIRECTIONAL MUTATION PRESSURES AND THE COMPOSITION OF MITOCHONDRIAL PROTEINS

Directional mutation pressures are implicated in most of the biases in animal mtDNA synonymous codon usage. Strong mutational pressures at the DNA level can also affect the amino acid composition of proteins (Sueoka 1988). The equilibrium base composition of a group of sites, such as first codon positions, will be determined by a mutation-selection-drift balance (Bulmer 1991). When mutational pressures are sufficiently strong to change the base composition of first codon positions, the frequency of amino acids encoded by these positions will change as well.

Introduction

The theory of directional mutation pressure (Sueoka 1962, 1988, 1995) is essentially an extension of the neutral mutation theory (Kimura 1983) and was developed to explain variation in *GC* content among bacterial genomes. When there is no selection and no directional mutation pressure, the rate of substitution from *GC* base pairs to *AT* base pairs is equal to the rate of substitution from *AT* base pairs to *GC* base pairs, and the equilibrium base composition is 50% *GC* and 50% *AT*. When these rates are unequal, the equilibrium base composition shifts.

Directional Mutation Pressure Theory

If p is the *GC* content of a sequence, $(1-p)$ is the *AT* content, μ is the rate of change from *GC* to *AT* base pairs and ν is the rate of change from *AT* to *GC* base pairs, then the equilibrium *GC* content is $\hat{p} = \nu / (\mu + \nu)$. The mutational pressure, μ_D , is defined as $\nu / (\mu + \nu)$, which is equal to 0.5 in the absence of directional pressure, greater than 0.5 if the mutational pressure favors *GC* base pairs and less than 0.5 if the mutational pressure favors *AT* base pairs.

Derivation from Sueoka (1988):

$$\text{change in } p \text{ in one generation: } \Delta p = \nu(1 - p_t) - \mu p_t = \nu - (\mu + \nu)p_t$$

at equilibrium:

$$\Delta p = 0, \quad p_t = p_{t+1} = \hat{p}$$

and

$$0 = \nu - (\mu + \nu)\hat{p} \quad \text{or} \quad \hat{p} = \nu / (\mu + \nu)$$

The mutational pressure, μ_D , can be estimated by the equilibrium *GC* content of sites free from selective constraints. Sueoka (1988) chooses to use P_3 , the composition of third codon positions. Assuming the neutrality of P_3 and an equilibrium between directional mutation pressure and selective constraints at other positions in DNA sequences, the effect of mutational pressures on these positions can be quantified. Plots of P_1 , P_2 , or P_{12} vs P_3 , where the subscript identifies codon position, show a linear relationship in which the frequency of *GC* base pairs at first and/or second positions increases with increasing μ_D .

With a mutation-selection equilibrium the relationship can be described by:

$$P_{12} = E_p + \epsilon_{12}(P_3 - E_p) + \text{error} \quad \text{or} \quad P_{12} = \epsilon_{12}P_3 + (1 - \epsilon_{12})E_p + \text{error}$$

where ϵ_{12} is the regression coefficient (slope of P_{12} vs. P_3) and E_p is the intercept where $P_{12}=P_3$. Note that this is based on the equation of a line given the slope and one point : $y - y_1 = m(x - x_1)$. In the absence of selection, the slope of this line should be equal to 1 and with complete selective constraint, the slope should be equal to 0. Thus, Sueoka (1988) uses ϵ_{12} as a “convenient” measure of neutrality.

Strand-Specific Directional Mutation Pressures

Sueoka's (1988) approach to directional mutation pressures is based on a conceptual model of equal mutational spectra for the two strands of DNA. When there is no strand-bias to the mutational process, the frequency of *A* equals the frequency of *T*, and *G* equals *C* irrespective of *GC* content (Sueoka 1995). Violations of this rule implicate the action of either selection or strand-specific mutational biases (Sueoka 1995). When the pattern of mutation differs between strands, as is clearly the case for mtDNA, the theory of directional mutation pressure describes only one aspect of the mutational bias.

Consider a sequence that is 50% *GC* base pairs at sites experiencing no selection. If these sites are 5% *G* and 45% *C*, the sequence is experiencing a directional mutation pressure that the theory does not detect. The base composition of fourfold degenerate sites from human mtDNA is quite similar to this example, and it is the *GC*-skew reported in Chapter II that is overlooked by this theory. An additional implication is illustrated by the nematode composition from Chapter II. *Ascaris suum* and *Caenorhabditis elegans* fourfold degenerate sites are both approximately 85% *AT* base pairs. The detected mutational pressure is equal, but the variation in *AT*-skew indicates that the pattern of mutation is different for these two species.

The inadequacy of μ_D as a measure of directional mutation pressures in mitochondrial

DNA is likely to confound attempts to describe the effect of mutational biases on the evolution of mitochondrial proteins. For example, a high frequency of glycine, which is encoded by the triplet *GGG*, may be correlated with a $\mu_D=0.7$ favoring *GC* base pairs.

This appears to be evidence that mutational pressure elevates the frequency of *G* at the first and second codon position. However, if the frequency of *C* is 0.65 and the frequency of *G* is 0.05, natural selection may be a more likely explanation for the high glycine content.

Mitochondrial Mutation Pressures

There are relatively few studies of the effect mutation pressures have on mitochondrial proteins. Jukes and Bhushan (1986) report a positive correlation between *GC* content of synonymous and nonsynonymous sites from the 13 protein coding genes of the human, cow, mouse, frog and fruit fly mitochondrial genomes. They also compare amino acid composition of these proteins. Some, but not all, amino acids encoded with *A* and *T* in the first two codon positions are used in higher frequency in the fruit fly genome where the mutational pressure favors *AT* base pairs. Likewise, some amino acids encoded with *G* and *C* in the first two codon positions are used less often in the fruit fly genome. Jermini et al. (1994) use a minor modification of Sueoka's μ_D to survey mutational pressures in cytochrome *b* genes from 110 taxa. *GC* content of nonsynonymous sites is positively correlated with mutation pressure in this data as well. Both of these studies are subject to the limitations of μ_D discussed above. Only Asakawa et al. (1991) consider strand-specific mutation pressures in any depth. They compare the base composition of first, second and third codon positions of genes encoded on one strand in each of 8 deuterostome taxa and contrast this composition with that of genes encoded on the other strand. For each taxon, the base composition of third codon positions from one strand is nearly a mirror image of the base composition of third codon positions from the other strand. Within each strand,

the base compositional bias of first and second codon positions is similar, though less extreme than the pattern of bias at third codon positions.

This chapter presents a more extensive study of the effects of mutation pressures on the composition of mitochondrial proteins using data from all published complete metazoan mitochondrial genomes. Following the lead of Asakawa et al. (1991) each base is considered separately to accommodate the strand-specific mitochondrial mutational biases. This study first examines the utility of fourfold degenerate third codon position composition as a measure of mutational spectra, then explores correlations of composition at first and second codon positions with mutation pressures. Unlike all three previous analyses described above, first and second codon positions are treated separately because of proposed differences in the selective constraints acting on these positions (Naylor et al. 1995). To complement the only study of mutational pressure in a single gene (Jermin et al. 1994), this analysis explores variation in response to mutation pressures among the 13 mitochondrial genes. This study concludes with further consideration of the effect of mutation pressure on the frequency of individual amino acids begun by Jukes and Bhushan (1986).

Methods

I tabulated the frequencies and relative frequencies of the four bases *G*, *A*, *T*, and *C* at each of the three codon positions and at fourfold degenerate codon positions, codon usage and predicted amino acid usage from each protein coding gene of thirty-one complete animal mitochondrial genomes. The 31 sequences listed in Table 4.1 constitute all complete metazoan animal mitochondrial genomes in the GenBank database as of November, 1995 when this analysis was initiated. This data set includes 19 vertebrate taxa, 2 echinoderms, 4 arthropods, 3 molluscs, 2 nematodes and an annelid. Both nematode genomes lack an open reading frame corresponding to the *Atpase 6* gene found in other taxa. No NADH subunit 4 is reported for *Artemia*. All other taxa have the standard

complement of 13 protein coding genes. Consequently, compositional data were collected for a total of 400 mitochondrial genes.

The frequency data were assembled with the assistance of a PASCAL program called USAGE (Appendix A). USAGE requires two input files for each genome sequence. One is a text file containing only the complete sequence. The other is a text file specifying name, beginning and ending position, and polarity of each protein coding gene. A sample input file is provided in Appendix B. USAGE generates three output files. The standard output file is generated as the program runs, producing a log of the sequence file name, the data range file name, and the sequence corresponding to the beginning and end of each reading frame. A sample output file is provided in Appendix C. This log allows the user to check whether these sequence ends correspond to appropriate start and stop codons as well as indicating when the data range specified is not a multiple of three. This situation arises occasionally in mitochondrial genomes when partial stop codons in the DNA sequence are completed by post-transcriptional polyadenylation. The COMP output file contains the frequency of *G*, *A*, *T* and *C* at the first, second, third and fourfold degenerate third codon positions as well as the frequency of each amino acid in the predicted peptide sequences. The CODON output file contains a complete codon usage table for each gene specified in the range file. Individual COMP and CODON files from each taxon were compiled into one large text file for use in other applications. Simple linear regressions were fit using SAS v6.0 and all graphs were generated using Cricket Graph v1.5.3.

Table 5.1. Taxa, GenBank accession numbers and citations for 31 complete metazoan mitochondrial genomes.

Taxon	Accession Number	Reference
<i>Albinaria coerulea</i>	X83390	Hatzoglou et al. 1995
<i>Anopheles gambiae</i>	L20934	Beard et al. 1993
<i>Apis mellifera</i>	L06178	Crozier and Crozier 1993
<i>Artemia franciscana</i>	X69067	Perez et al. 1994
<i>Ascaris suum</i>	X54253	Okimoto et al. 1992
<i>Balaenoptera musculus</i>	X72204	Arnason and Gullberg 1993
<i>Balaenoptera physalus</i>	X61145	Arnason et al. 1991
<i>Bos taurus</i>	V00654	Anderson et al. 1982
<i>Caenorhabditis elegans</i>	X54252	Okimoto et al. 1992
<i>Cepaea nemoralis</i>	U23045	Terrett et al. 1995
<i>Crossostoma lacustre</i>	M91245	Tzeng et al. 1992
<i>Cyprinus carpio</i>	X61010	Chang et al. 1994
<i>Didelphis virginiana</i>	Z29573	Janke et al. 1994
<i>Drosophila yakuba</i>	X03240	Clary and Wolstenholme 1985
<i>Equus caballus</i>	X79547	Xu and Arnason 1994
<i>Gallus gallus</i>	X52392	Desjardins and Morais 1990
<i>Gorilla gorilla</i>	D38114	Horai et al. 1995
<i>Halichoerus grypus</i>	X72004	Arnason et al. 1993
<i>Homo sapiens</i>	V00662	Anderson et al. 1981
<i>Katharina tunicata</i>	U09810	Boore and Brown 1994
<i>Lumbricus terrestris</i>	U24570	Boore and Brown 1995
<i>Mus musculus</i>	V00711	Bibb et al. 1981
<i>Oncorhynchus mykiss</i>	L29771	Zardoya et al. 1995
<i>Pan paniscus</i>	D38116	Horai et al. 1995
<i>Pan troglodytes</i>	D38113	Horai et al. 1995
<i>Paracentrotus lividus</i>	J04815	Cantatore et al. 1989
<i>Petromyzon marinus</i>	U11880	Lee and Kocher 1995
<i>Phoca vitulina</i>	X63726	Arnason and Johnsson 1992
<i>Pongo pygmaeus</i>	D38115	Horai et al. 1995
<i>Rattus norvegicus</i>	X14848	Gadaleta et al. 1989
<i>Strongylocentrotus purpuratus</i>	X12631	Jacobs et al. 1988

Results

Mutational Pressures at Fourfold Degenerate Sites

Fourfold degenerate third codon positions may provide the best possible estimate of the equilibrium composition generated by mitochondrial mutation pressures, yet the loglinear analyses presented in Chapter III indicate that synonymous site base composition varies among fourfold degenerate codon families. At present, it is not possible to exclude selection as a mechanism contributing to this variation. Therefore, it is important to examine the nature and magnitude of variation among codon families before using overall fourfold degenerate site composition as an estimate of mutational pressures. In Figures 5.1-5.4 third position base composition for each fourfold degenerate codon family is plotted against the average fourfold degenerate site composition. Each of the 400 points on a plot corresponds to the proportion of codons ending in a particular nucleotide in a single gene from a single taxon.

1. The diagonal line on each plot represents the expectation of complete concordance between the third position base composition for each codon family and the average composition of fourfold degenerate sites. Most codon family-base comparisons show general agreement with this expectation. That is, all comparisons except serine codons ending in C show a strong positive correlation along, or near, the predicted line.
2. All plots show some variance about the predicted line, and in some plots, most notably arginine codons ending in A, this variance can be quite large. Arginine is the least frequent of the amino acids encoded by a fourfold degenerate codon family, and it is likely that this variance is due to error associated with small sample size. Likewise, many plots show a number of points at either extreme (0 or 1) of the y-axis. These points generally correspond to estimates from small genes.

Figure 5.1. Relative frequency of each base at third positions of leucine (column 1) and valine (column 2) codons vs. relative frequency of the base at all fourfold degenerate sites.

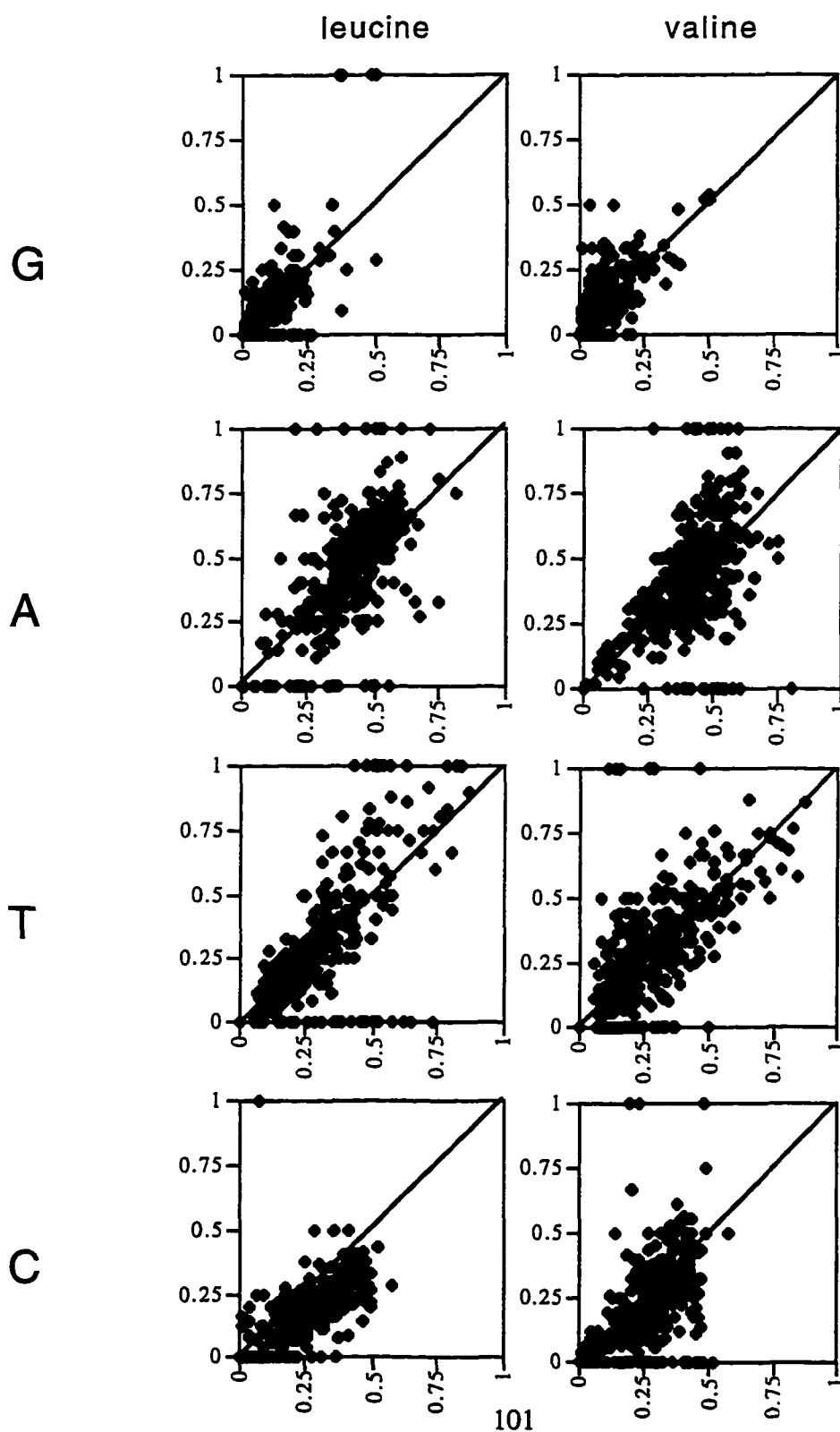


Figure 5.2. Relative frequency of each base at third positions of arg (column 1) and gly (column 2) codons vs. relative frequency of the base at all fourfold degenerate sites.

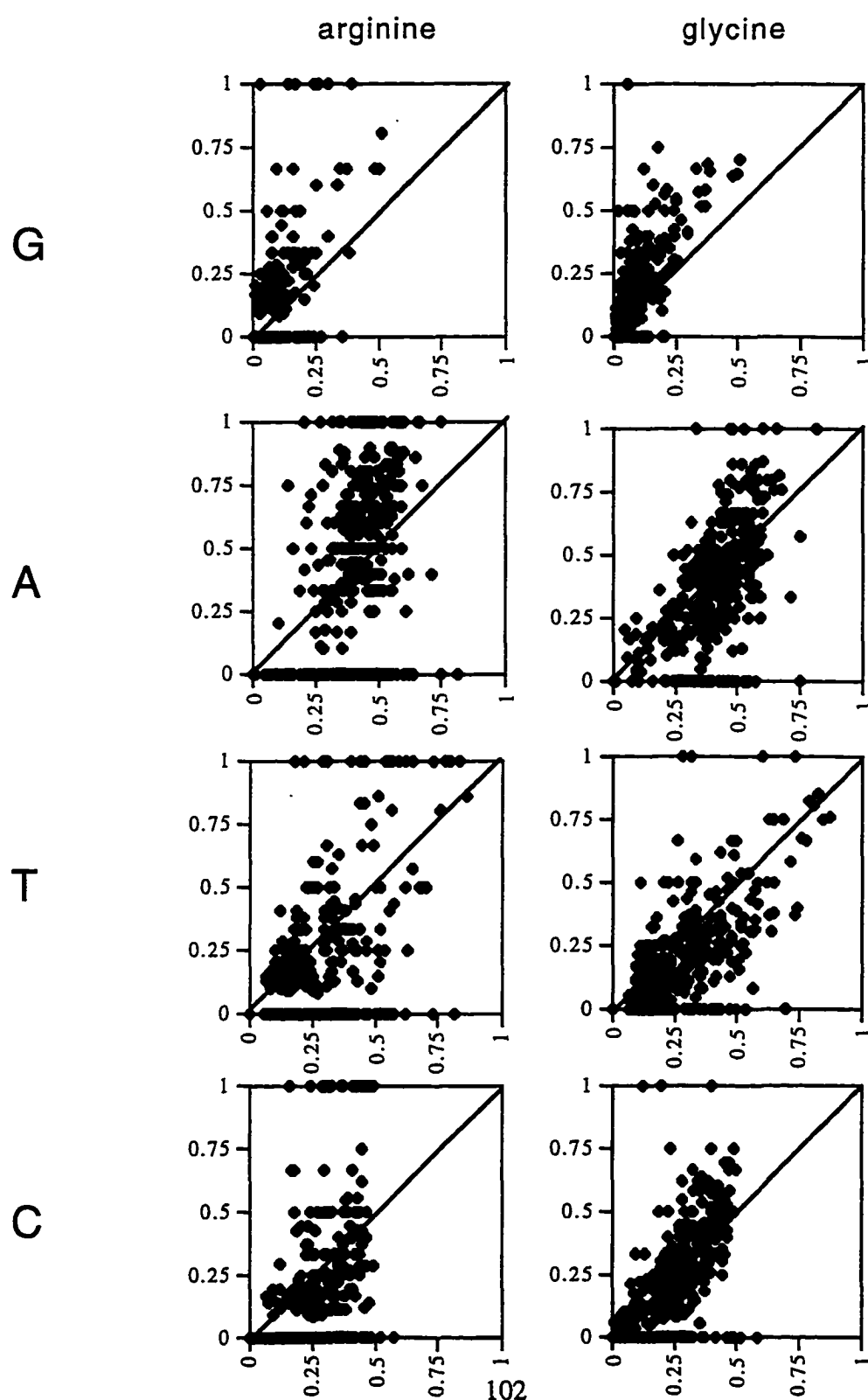


Figure 5.3. Relative frequency of each base at third positions of ser (column 1) and thr (column 2) codons vs. relative frequency of the base at all fourfold degenerate sites.

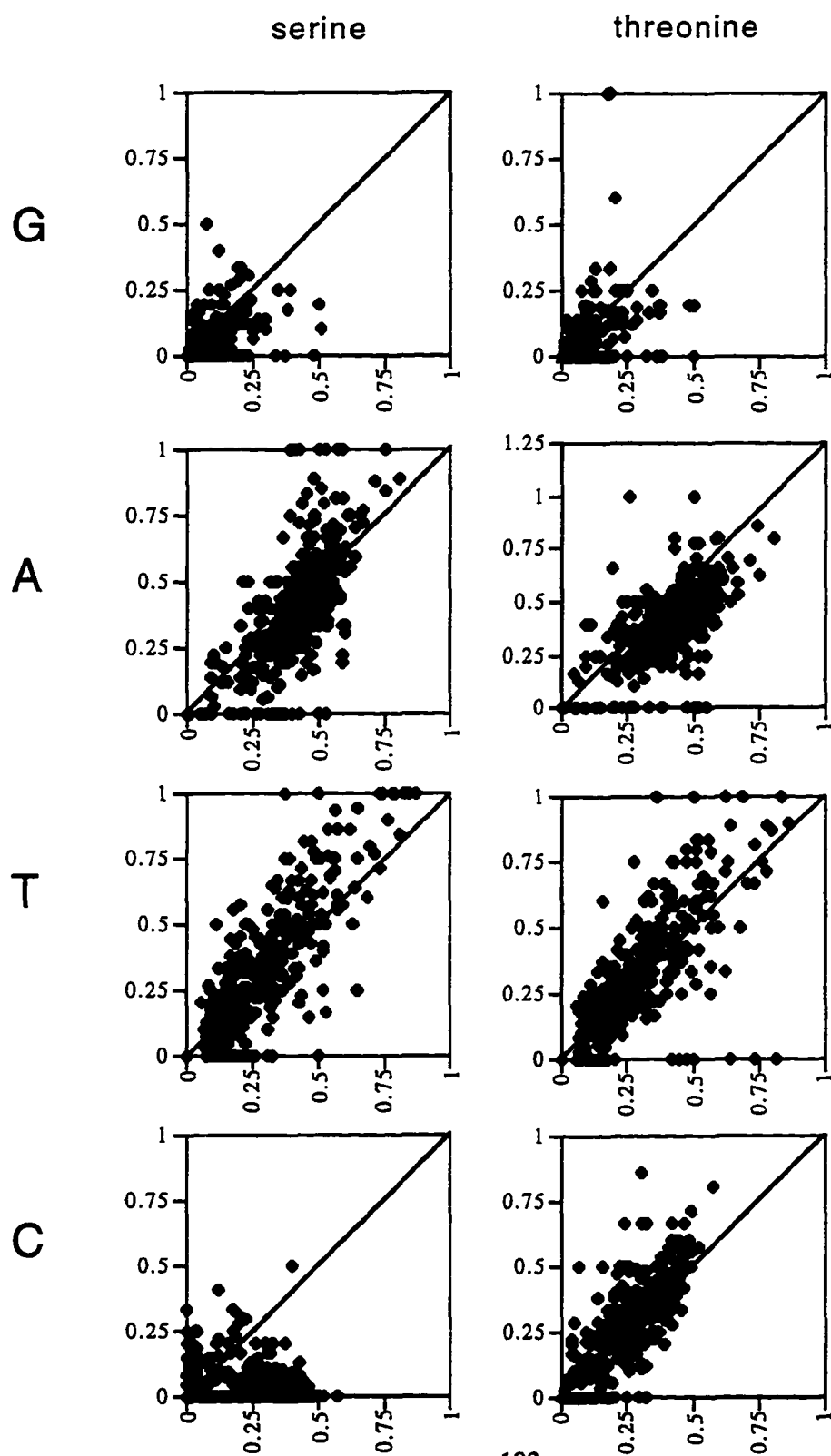
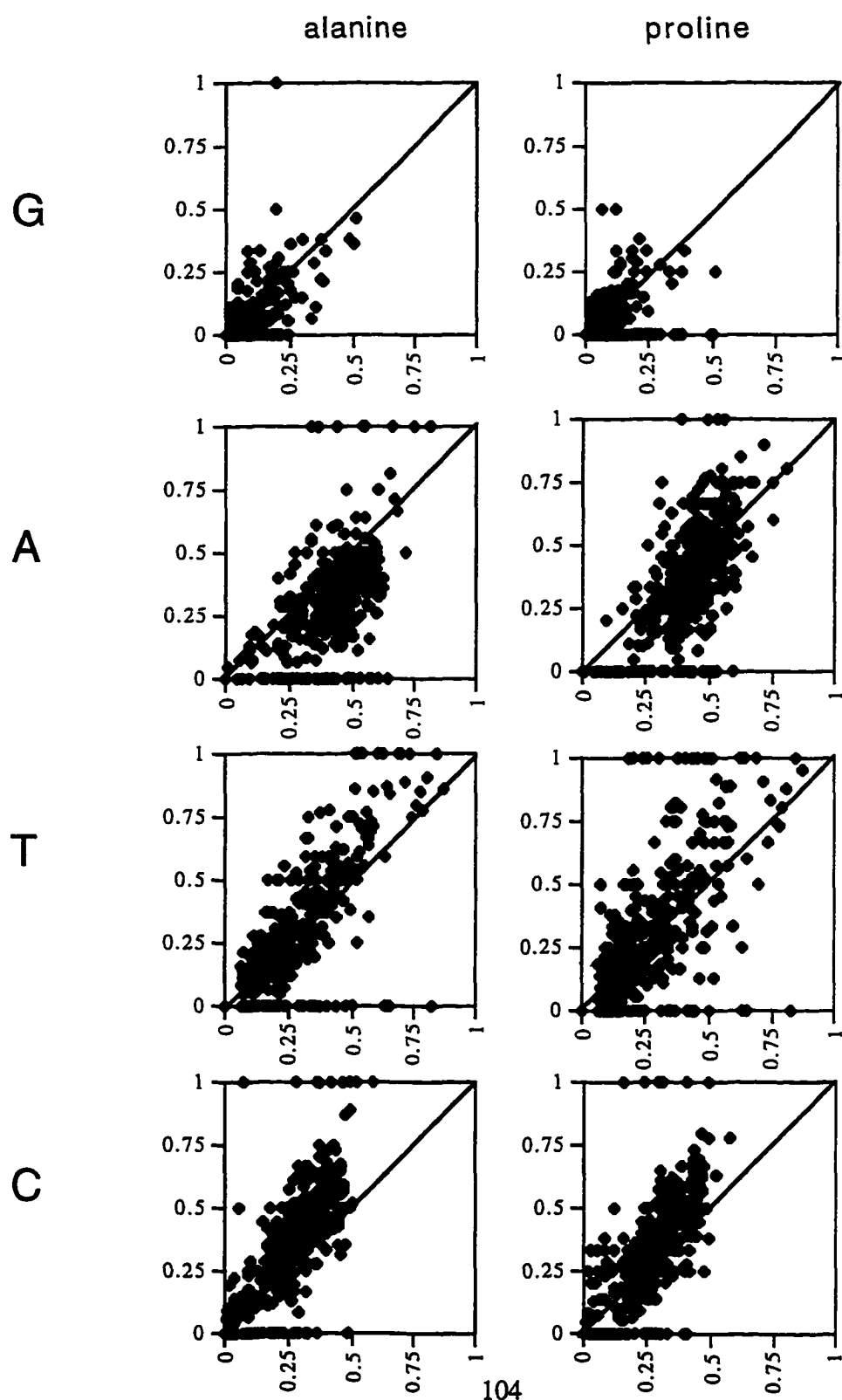


Figure 5.4. Relative frequency of each base at third positions of ala (column 1) and pro (column 2) codons vs. relative frequency of the base at all fourfold degenerate sites.



3. Despite the overall agreement with expectations, there are apparent compositional differences between some codon families. Leucine codons end in *A* more often and *C* less often than average fourfold degenerate codons. Serine codons rarely end in *C*. In contrast with leucine, alanine codons end in *C* more often and *A* less often than average, although plots for *G* and *T* are similar for both amino acids. A high frequency of codons ending in *G* is only observed in the arginine and glycine codon families.

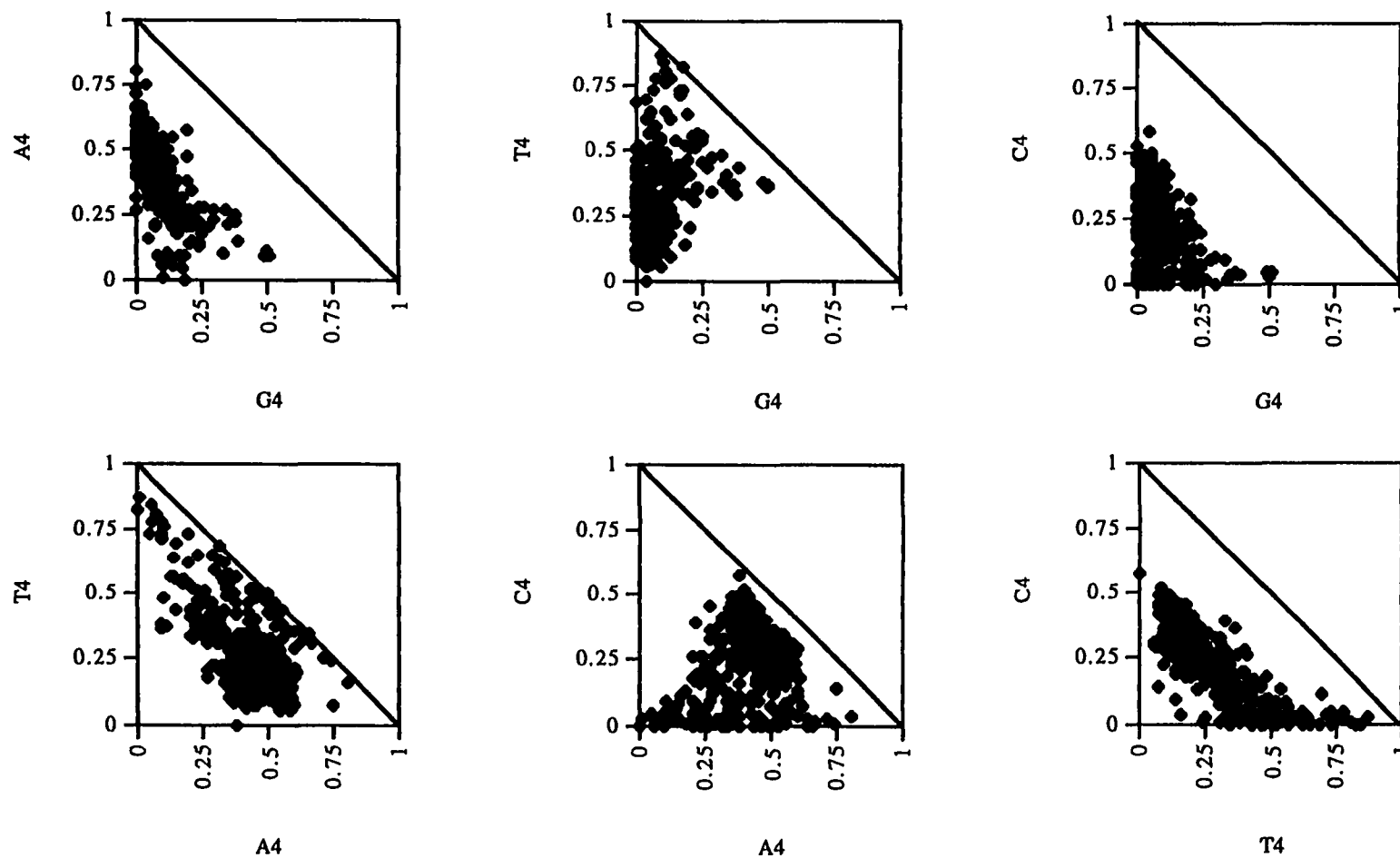
4. Previous analysis indicated that codon families with the same base at the second codon position have similar composition at the third position. This observation appears to be consistent with the plots corresponding to leucine and valine codons (second position *T*), as well as the plots corresponding to arginine and glycine codons (second position *G*). The remaining four codon families, serine, proline, threonine and alanine, all have *C* at the second codon position. Plots for codons corresponding to these amino acids are also similar, with the exception of serine codons ending in *C*. The low frequency of these serine codons is unlike the pattern of usage in any other codon family.

Overall, the variability among codon families appears limited enough to justify the assumption that fourfold degenerate sites are a reasonable estimate of mitochondrial mutation pressures. This assumption, and its less restrictive relative that includes all third codon positions in such estimates, certainly have historical precedent. Although this error in the estimates of mutational pressure seems unlikely to produce false correlations in the remainder of this analysis, it does suggest that some caution should be exercised in interpretation of quantifications of the effect of mutational pressures on other sites in the genome.

Strand-specific mutational pressures require that compositional correlations be addressed with each of the four bases individually. The relative frequencies of these four

bases must sum to one, so the four individual measures of mutational pressures are not independent. Furthermore, specific mutational pathways link the frequencies of the two bases involved and can lead to patterns among the four measures. For instance, no animal mitochondrial genome is known in which mutational pressures simultaneously favors both *C* and *T* at silent sites on the same strand. Figure 5.5 illustrates the relationship between all six combinations of two bases. *GC* content never exceeds approximately 65% and conversely, *AT* content never falls below about 35%. Plots of *A* vs. *G*, *C* vs. *T*, and *T* vs. *A*, show clear negative correlations. The remaining plots show a more complicated relationship. Low frequencies of *T* are never associated with the highest frequencies of *G*. Low frequencies of *A* are never associated with the highest frequencies of *C*. The observed relationships between *A* and *T* and between *G* and *C* underscore the necessity of treating these bases separately in analyses of the effects of mitochondrial mutation pressures on nonsynonymous sites.

Figure 5.5. Plots of all pairwise comparisons of the relative frequencies of each of the four bases at fourfold degenerate third codon positions. No point can exceed the diagonal line where the sum of the two frequencies equals one.



Mutational Pressures at First and Second Codon Positions

Figures 5.6 and 5.7 show a positive correlation between usage of each base at the first and second codon positions, respectively, and the relative frequency of the base at fourfold degenerate sites. Black diamonds represent genes with mRNA's of the same polarity as the published complete genome sequence. White diamonds represent genes encoded on the opposite strand, limited to ND6 from deuterostome taxa, four genes from each insect taxon, *Albinaria* and *Cepaea*, six from *Katharina*, and two from *Artemia*. Replication mechanisms for the molluscan genomes have not been characterized, which precludes unambiguous identification of which strands are "homologous" between some taxa. However, the data for the two strands are largely overlapping and justify treating both strands in the same analysis. One possible exception is observed in the plots corresponding to *G* at the first and second codon position, where vertebrate nd6 genes exhibit the highest frequency of *G* at fourfold degenerate sites of any sequence analyzed. Simple linear regressions of the relative frequency of each base at first codon positions using the relative frequency of the base at fourfold degenerate sites as a predictor are highly significant for all four bases ($p < 0.0001$). Similar regressions of second codon position base composition are also highly significant (Table 5.2). Residual analysis indicated that vertebrate ND6 points are highly influential in regressions for *G* usage. Removing ND6 from the data has only a modest effect on the estimated slope and definitely does not eliminate the strong correlation between composition of first or second codon positions with fourfold degenerate site composition.

Figure 5.6. Relative frequency of each base at all first codon positions vs. relative frequency of the base at all fourfold degenerate third codon positions for 400 mitochondrial genes from 31 taxa. Black diamonds represent genes with mRNA's of the same polarity as the published complete genome sequence. White diamonds represent genes encoded on the opposite strand.

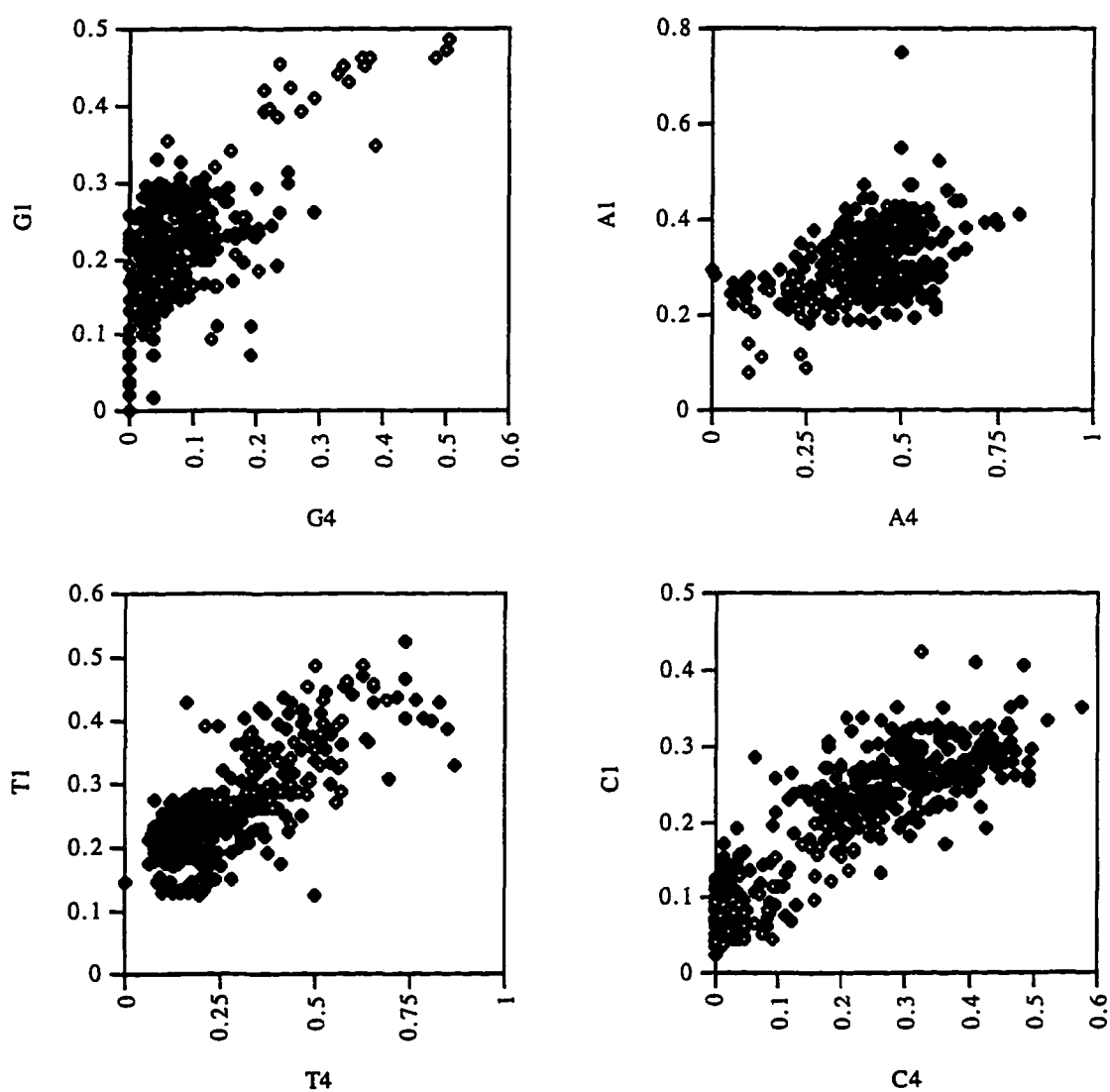


Figure 5.7. Relative frequency of each base at all second codon positions vs. relative frequency of the base at all fourfold degenerate third codon positions for 400 mitochondrial genes from 31 taxa. Black diamonds represent genes with mRNA's of the same polarity as the published complete genome sequence. White diamonds represent genes encoded on the opposite strand.

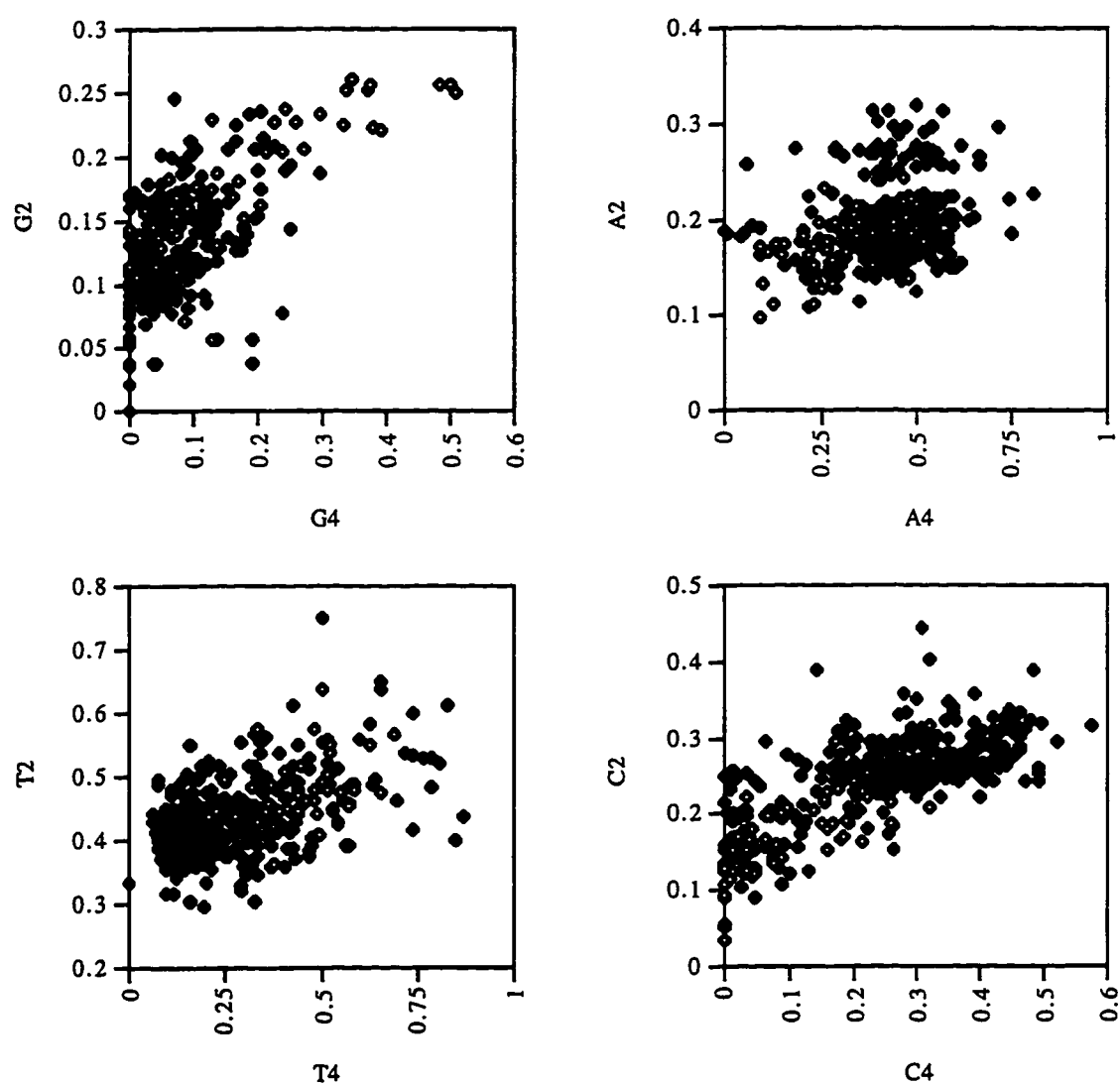


Table 5.2. Summary output from simple linear regressions of the relative frequency of a base at the first (G1, A1, T1, C1) or second (G2, A2, T2, C2) codon positions and the relative frequency at fourfold degenerate sites.

Base	Predictor	F	$p>F$	$adj-R^2$	intercept	slope
G1	G4	282.525	0.0001	0.4152	0.1670	0.6249
G2	G4	262.259	0.0001	0.3972	0.1040	0.3530
A1	A4	80.935	0.0001	0.1690	0.2111	0.2241
A2	A4	34.971	0.0001	0.0808	0.1582	0.0851
T1	T4	587.318	0.0001	0.5961	0.1640	0.3712
T2	T4	127.275	0.0001	0.2423	0.3880	0.1750
C1	C4	911.599	0.0001	0.6961	0.0990	0.4950
C2	C4	489.943	0.0001	0.5518	0.1620	0.3298

If nucleotide usage at first and second codon positions was determined entirely by the equilibrium base composition of the mutational spectrum, and fourfold degenerate site composition was a good estimate of this equilibrium, the regressions shown in Table 5.2 would have a slope equal to one and intercept equal to zero. Deviations from this state are expected when natural selection constrains first and second position base composition. No estimated slope coefficient exceeds 0.6249 (first position G) indicating that selection constrains the use of all four bases. For every base, the slope is greater for the first codon position regression than for the second codon position regression suggesting that selection is stronger for the second base of the codon. This is consistent with the tendency of the genetic code to preserve the general character of amino acids within groups defined by the second codon base. R^2 values in Table 5.2 indicate the proportion of variation in nucleotide usage at first and second positions that can be explained by variation in mutational pressures reflected at fourfold degenerate sites. The R^2 values range from 0.0808 for second position A's to 0.6961 for first position C's, suggesting that mutational pressures could account for a significant proportion of variation in usage of at least some

amino acids among animal mitochondrial genomes.

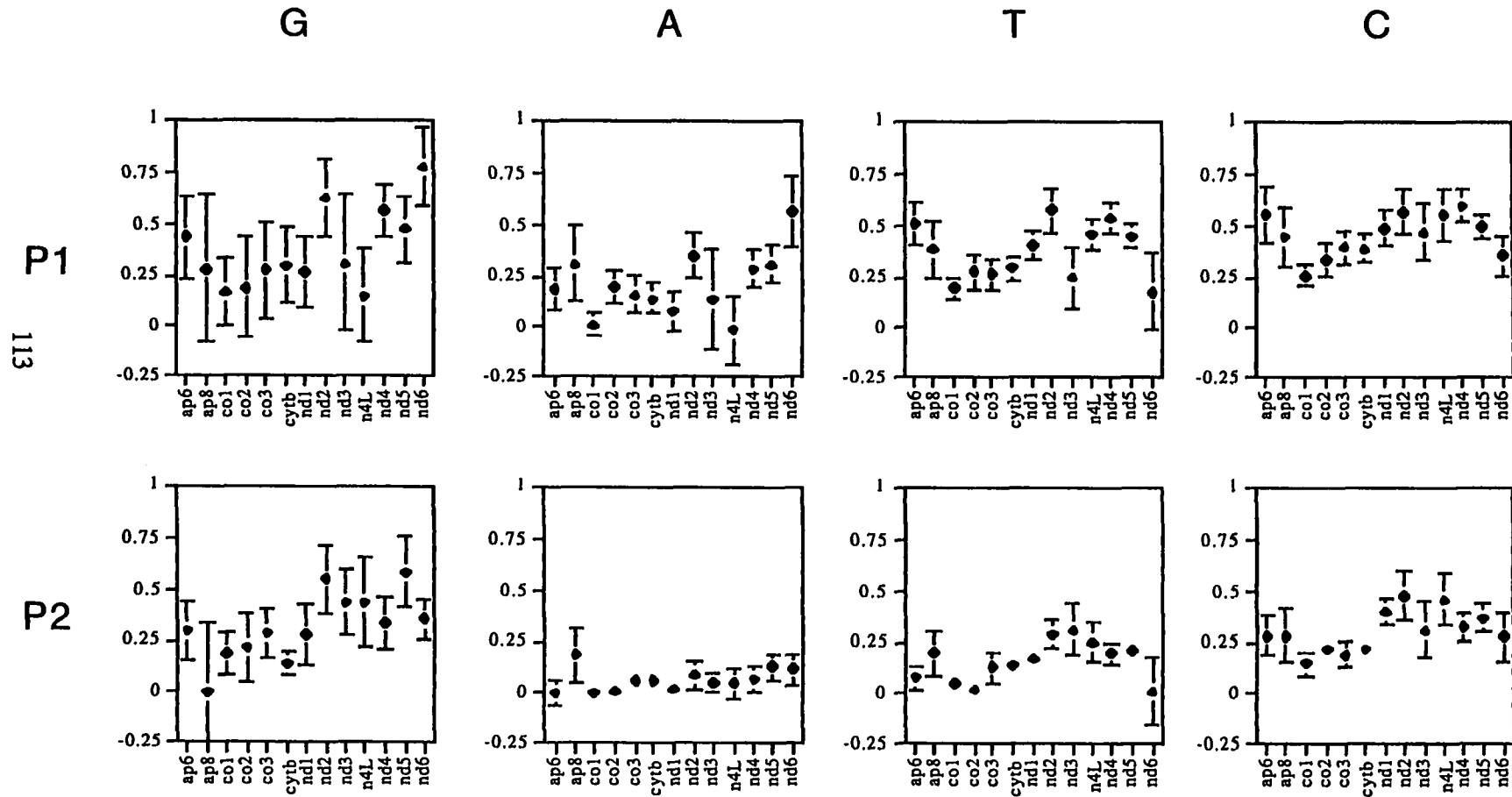
The logic behind measuring fourfold degenerate site composition separately for each gene is based on the idea that mutational pressures are known to vary within a genome. However, the benefit of accommodating intramolecular variation may be outweighed by the additional variation introduced by the small sample size for each gene, especially for low frequency bases in highly biased genomes. Average fourfold degenerate site composition, determined by summing the frequencies across all genes in a genome, was substituted for the gene by gene estimates of mutational pressures, and the regressions were refit using this new independent variable. There is very little difference in the relationships between first and second position composition and either measure of mutational pressures.

Variation in Response to Mutational Pressures among Non-homologous Genes

Differences in selective pressures between genes could result in differential response to mutational pressures. There is some evidence of variation in the rate of amino acid replacement between mitochondrial genes. I fit individual regressions of the relative frequency of each base at codon position 1 (and 2) against relative frequency at fourfold degenerate sites for each gene. Variation in selective pressures should result in variation in the slope of the regression lines. Confidence intervals at the 95% level (Figure 5.8) for the estimated slope coefficients illustrate several points.

1. The estimated slope coefficients are significantly different from zero for $85/104 = 82\%$ of the individual tests at the 0.05 level. In all significant regressions the slope is greater than zero. If we adjust the significance level to achieve an experiment wide error of 0.05 for 104 tests, the critical p -value drops to 0.0005. At this level $66/104 = 63\%$ of the slope coefficients are significantly different from zero and the confidence intervals widen considerably. It may be important to note that this does not mean that positive slopes are

Figure 5.8. 95% Confidence Intervals for the estimated slope of regressions of first (row 1) or second (row 2) codon position composition and fourfold degenerate sites for each gene.



not plausible for insignificant regressions, but rather indicates a lack of resolution for particular gene/base comparisons.

2. The standard errors for regressions of the frequency of *G* at position 1 (and 2) are larger than those estimated for any of the other three bases. While the frequency of *G* at first and second positions is only slightly lower than the frequencies of *A*, *T* and *C* at these positions, at fourfold degenerate sites, *G* occurs at very low frequency. The large regression errors may arise from error in estimating the mutational pressures from the base composition of fourfold degenerate sites. This problem would be aggravated by subdividing the data by gene, a cost offset by accommodating intramolecular variation in mutational pressures.

3. At the 0.05 level there are significant differences (non-overlapping confidence intervals) among the slope coefficients of individual genes. If the slope is a reasonable index of neutrality, this variation is indicative of selective differences among genes. Although there are a few exceptions, the observation that regressions for position 1 have greater slope than regressions for position 2 in the complete data, holds true for individual genes. Variation in slope for regressions of different bases are also still evident for many comparisons, indicating that the magnitude of selection against classes of amino acids grouped by first or second codon position varies. For instance, in most genes, selection appears to constrain the frequency of codons beginning with *A* to a greater extent than it constrains the frequency of codons beginning with *C*. Consequently, we observe greater variation in the frequency of leucine (*CTN*), proline (*CCN*), histidine (*CAY*), glutamine (*CAR*) and arginine (*CGN*) among homologous genes from diverse organisms than variation in the frequency of isoleucine (*ATY*), methionine (*ATR*), threonine (*ACN*), asparagine (*AAV*), lysine (*AAR*), serine (*AGY*) and arginine (*AGR*). Whether the variation in slope between

bases is due to the effect of selection on one, several or all amino acids within rows (position 1) or columns (position 2) of the genetic code, will be explored elsewhere.

4. It is easy to discern similarities in the relative slopes for individual genes both between regressions for different bases and regressions for the two codon positions. The trends are especially evident in comparisons of *T1*, *T2*, *C1* and *C2*. These trends may be indicative of overall selective differences among genes. One possible test of this hypothesis is to look for concordant differences in the rate of amino acid replacements among genes. A simple estimate of replacement rates is the pairwise difference between two sequences that are not so divergent that multiple replacements become problematic, but divergent enough to register enough differences for the comparison. Figure 5.9 shows scatterplots of the estimated slopes for each gene/base comparison vs. pairwise difference between rat and mouse sequences listed in table 5.1. The difference matrix was tabulated from the Lee and Kocher (1995) alignment of 11 mitochondrial protein coding genes from 11 vertebrate taxa. The positive correlations for all four bases from both first and second codon positions are consistent with the idea that the estimated slopes are proportional to the selective constraints on these genes.

Figure 5.9. Estimated slope for first and second codon position regressions vs. relative rate of replacement estimated by pairwise amino acid differences between the rat and mouse sequences from the Lee and Kocher (1995) alignment.

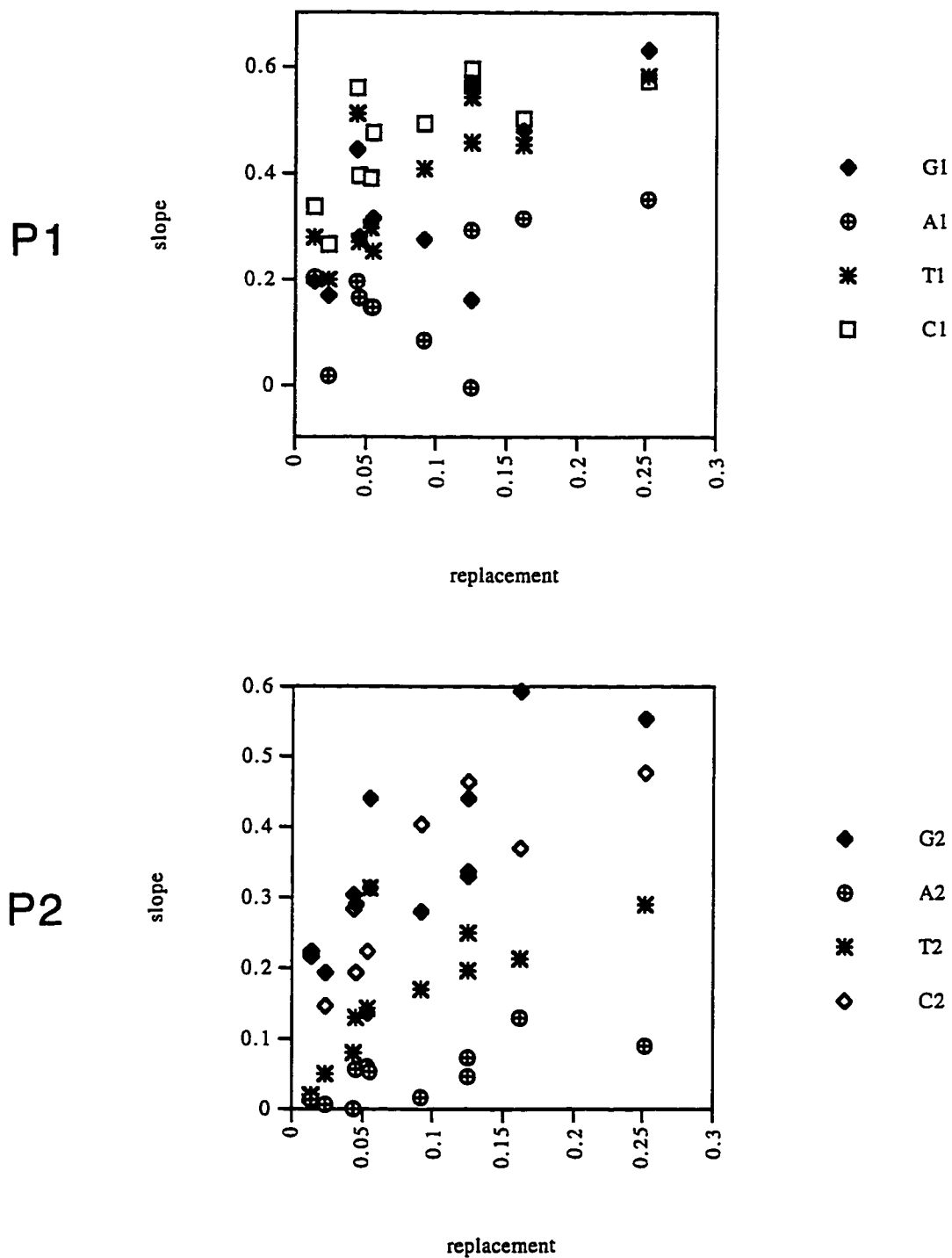


Table 5.3. Summary of 40 simple linear regressions of the relative frequency of an amino acid regressed on the relative frequency of a base (b4) found in either the first or second position of the codon for that amino acid.

amino acid	codon	b4	F	p>F	R ²	intercept	slope
I	ATY	T	0.001	0.9726	0.0000	0.0818	-0.0003
M	ATR	T	0.005	0.9440	0.0000	0.0588	-0.0005
A	GCN	G	0.502	0.4789	0.0013	0.0597	0.0120
W	TGR	G	0.577	0.4478	0.0014	0.0277	-0.0070
D	GAR	A	0.665	0.4154	0.0017	0.0281	-0.0048
W	TGR	T	0.910	0.3408	0.0023	0.0283	-0.0041
E	GAY	A	1.031	0.3106	0.0026	0.0200	-0.0052
R	CGN	G	4.159	0.0421	0.0103	0.0155	0.0111
S	TCN	C	5.958	0.0151	0.0147	0.0538	0.0168
K	AAR	A	6.023	0.0146	0.0149	0.0196	0.0190
R	CGN	C	6.975	0.0086	0.0172	0.0145	0.0079
E	GAY	G	7.184	0.0077	0.0177	0.0161	0.0224
S	TCN	T	8.002	0.0049	0.0197	0.0623	-0.0166
M	ATR	A	12.86	0.0004	0.0313	0.0452	0.0325
D	GAR	G	15.742	0.0001	0.0380	0.0233	0.0376
C	TGY	G	16.943	0.0001	0.0408	0.0079	0.0245
R	AGR	G	18.234	0.0001	0.0438	0.0063	0.0374
H	CAY	A	21.807	0.0001	0.0519	0.0103	0.0290
H	CAY	C	26.308	0.0001	0.0620	0.0156	0.0286
Y	TAY	A	26.519	0.0001	0.0625	0.0482	-0.0306
Q	CAR	A	27.912	0.0001	0.0655	0.0114	0.0265
R	AGR	A	36.217	0.0001	0.0834	0.0222	-0.0315
S	AGY	G	45.667	0.0001	0.1029	0.0117	0.0489
C	TGY	T	45.744	0.0001	0.1031	0.0048	0.0183
N	AAV	A	52.87	0.0001	0.1173	0.0188	0.0498
Q	CAR	C	54.909	0.0001	0.1212	0.0147	0.0325
I	ATY	A	69.01	0.0001	0.1478	0.0421	0.0961
Y	TAY	T	71.883	0.0001	0.1530	0.0254	0.0368
T	ACN	A	72.77	0.0001	0.1546	0.0318	0.0970
S	AGY	A	86.171	0.0001	0.1780	0.0316	-0.0392
A	GCN	C	96.505	0.0001	0.1952	0.0413	0.0823
P	CCN	C	143.743	0.0001	0.2653	0.0269	0.0947
V	GTN	T	154.794	0.0001	0.2800	0.0270	0.1091
F	TTY	T	203.361	0.0001	0.3382	0.0419	0.1122
G	GGN	G	235.862	0.0001	0.3721	0.0349	0.2392
T	ACN	C	240.649	0.0001	0.3768	0.0399	0.1367
V	GTN	G	411.789	0.0001	0.5085	0.0330	0.3134
L	CTN	T	587.763	0.0001	0.5963	0.1793	-0.2695
L	TTR	T	652.397	0.0001	0.6211	-0.0007	0.2237
L	CTN	C	752.681	0.0001	0.6541	0.0273	0.3314

Variation in Response to Mutational Pressures among Amino Acids

The variation in estimated slopes for different nucleotides indicates that individual amino acids are differentially affected by mutational pressures on the bases found in their codons. Table 5.3 summarizes the results of 40 separate simple linear regressions of the relative frequency of each amino acid and the relative frequency at fourfold degenerate sites of any base involved in the first or second position of the codon for that amino acid. For example, alanine (*GCN*) usage was regressed against the frequency of *G* at fourfold degenerate sites and against the frequency of *C* at fourfold degenerate sites. Leucine *UUY*, leucine *CUN*, serine *UCN*, serine *AGY*, arginine *CGN* and arginine *AGR* codons were treated separately. Several features of Table 5.3 are notable.

1. Most of the regressions are highly significant. Even with a Bonferoni adjustment for multiple comparisons to get an experiment wide $\alpha=0.05$, the critical p -value is 0.0013. Twenty-seven out of forty of these regressions are significant at this level. Of these, five show a significant negative slope coefficient contrary to the expectation of a positive correlation of amino acid composition and mutational pressures. This will be discussed later.

2. Among the twenty-two significant positive correlations, the variation in slope and R^2 shows considerable diversity in the effect of mutational pressures on protein composition. Setting aside leucine, which will be discussed separately, significant slopes range from 0.0183 (cysteine *TGY-T4*) to 0.3134 (valine *GTN-G1*). Significant regressions explain anywhere from 47% (valine *GTN-G4*) to just 3% (methionine *ATR-A4*) of the variation in usage of an amino acid among the total set of mitochondrial genes from 31 taxa. Histograms of the slopes and R^2 are shown in Figure 5.10. For most amino acids, the slope is less than 0.1 and the proportion of variation explained by the regression is less

than 0.2. Three amino acids with large slope and R^2 values, phenylalanine (*TTY*), glycine (*GGN*), and proline (*CCC*) are encoded by the same base at the first and second codon positions. Valine (*GTN*) usage shows a very strong regression with the frequencies of both *G* and *T* at fourfold degenerate sites.

3. The regressions shown in Table 5.3 are listed in order of increasing significance. Among the top of the list are regressions for glutamic acid, aspartic acid, lysine and arginine (*CGN*). It is not surprising that usage of these charged amino acids may be more constrained by selection in the largely hydrophobic membrane spanning mitochondrial proteins. However, regressions for histidine and arginine (*AGR*) are significant suggesting that mutational pressures may be strong enough to affect the frequency of some charged amino acids.

4. For most amino acids, the regression with the fourfold degenerate site base matching the first codon position is stronger than that for the base matching the second codon position. This general trend is consistent with the variation between slopes for overall first and second position regressions. Threonine (*ACN*) and alanine (*GCN*) are the two exceptions and the strong correlations of these amino acids with the mutational pressure affecting *C* contribute to the large slope for *C2* in Table 5.2.

5. Figure 5.11 provides a visual representation of the relative slope and significance of each individual regression from Table 5.3 in the familiar context of the genetic code table. Some amino acids have a much greater effect on the overall first and second position regressions than other amino acids. For example, valine (*GTN*) and glycine (*GGN*) exhibit a much larger slope than other amino acids encoded by *G* at the first position. Leucine (*TTR*) and phenylalanine (*TTY*) drive the first position correlation with *T*.

Figure 5.10. Histograms of estimated slope and R^2 values from the regressions of individual amino acid frequencies from Table 5.3. Amino acids and the fourfold degenerate base corresponding to the highest values are identified above the frequency bars.

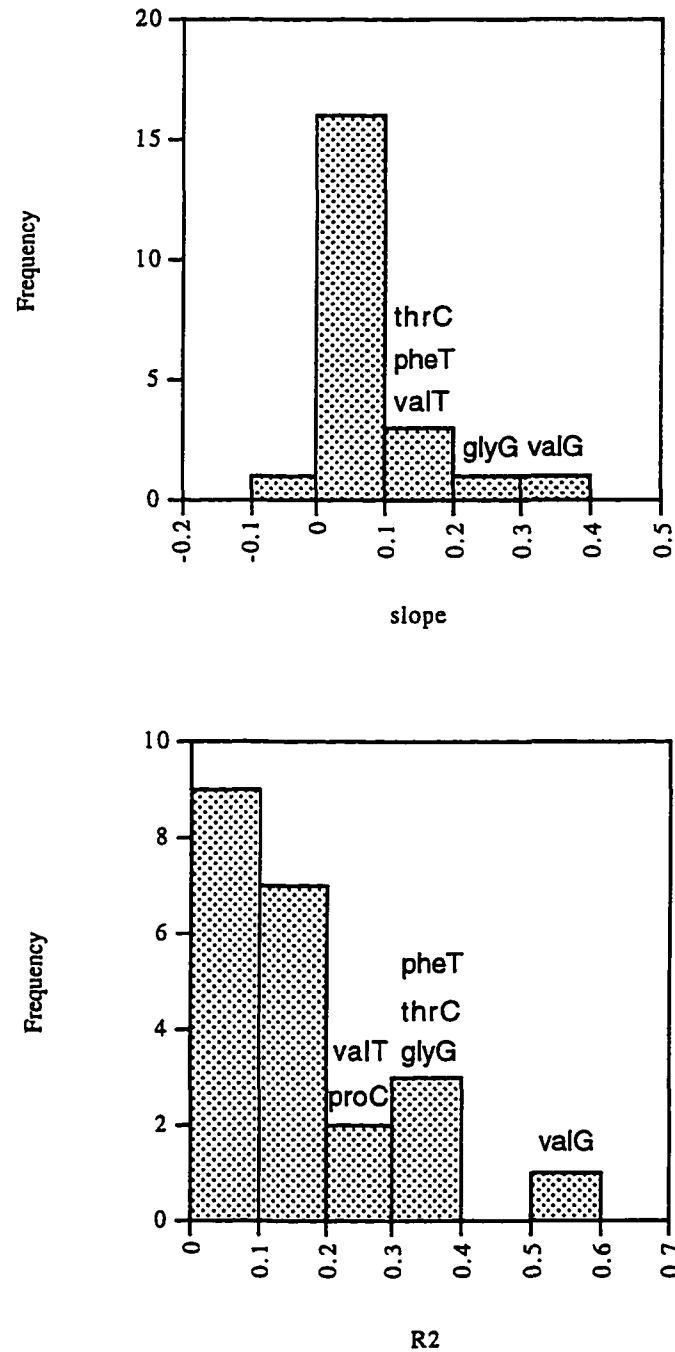
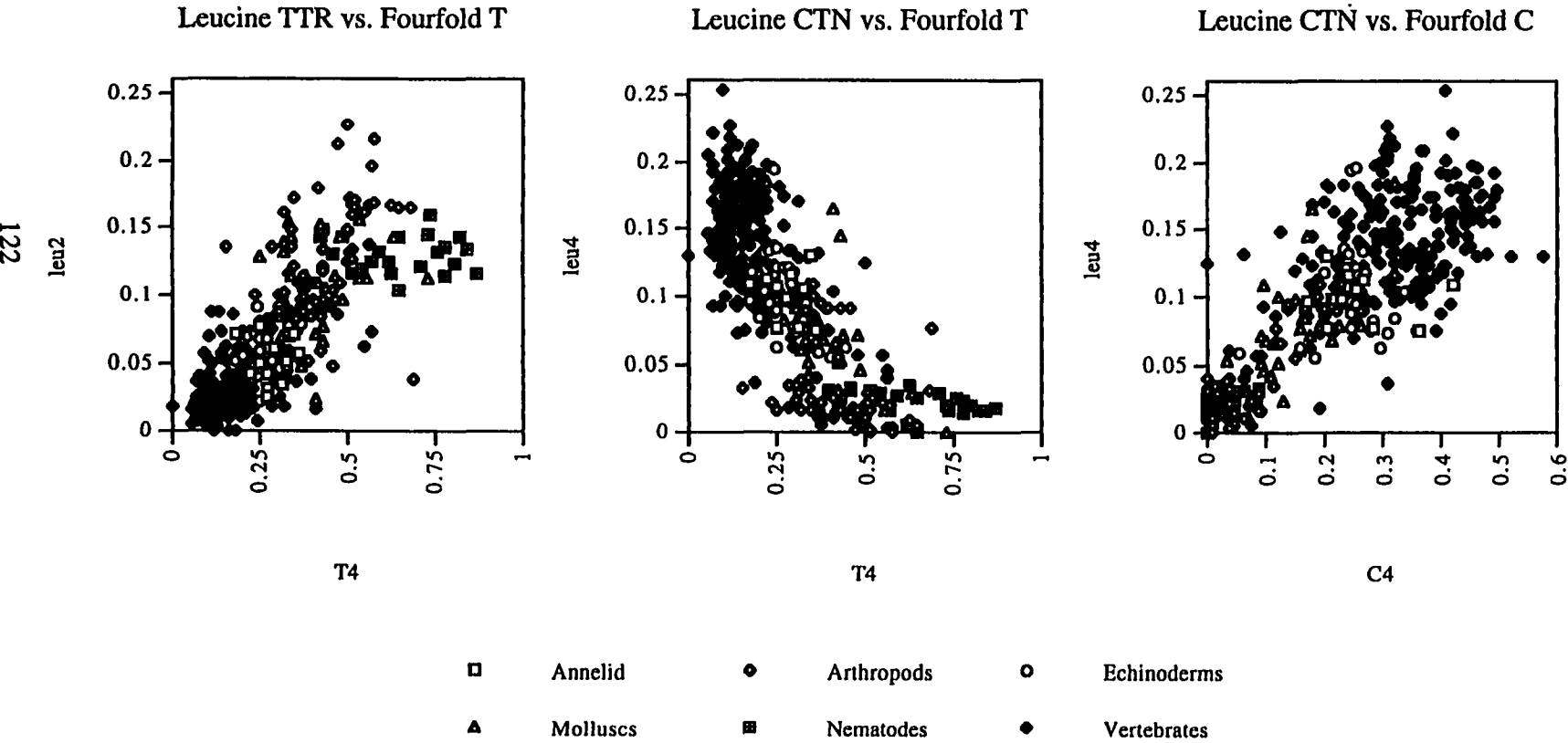


Figure 5.11. A visual representation of the relative slope and significance level of each individual regression from table 5.3. The size of each first or second position base indicates the magnitude of the estimated slope. Bold letters indicate that the slope is significant at the 0.05 level. Outlined letters indicate that the slope is significantly less than zero.

		Second Codon Position			
		T	C	A	G
First Codon Position	T	TT_Y TT_R	TCN	TAY	TGY TGR
	C	CT_N	CC_N	CAY CAR	CGN
	A	AT_Y ATR	AC_N	AA_Y AAR	AG_Y AGR
	G	GT_N	GC_N	GAY GAR	GG_N

6. The three strongest regressions in Table 5.3 are all related to leucine usage, *CTN-C4*, *CTN-T4*, and *TTR-T4*, yet there is little overall correlation between total leucine usage and the relative frequency of either *T* or *C* at fourfold degenerate sites. The average mitochondrial protein is about 17% leucine. The highly significant slopes exist because mutational pressures strongly influence whether the *TTR* or *CTN* codons predominate. Figure 5.12 shows that in taxa like nematodes and arthropods, where *T* is strongly favored by the mutational spectrum, the *TTR* codons account for most of the leucines found in

Figure 5.12. Shifts in usage of twofold (*TTR*) and fourfold (*CTN*) degenerate codons for leucine among diverse taxa.



these proteins, whereas in vertebrates, where most genes are encoded on a strand where mutational pressures favor *C* over *T*, the *CTN* leucine codons are more abundant. This accounts for the strong negative correlation between *CTN* leucines and fourfold degenerate *T* frequency.

7. There are three other highly significant negative slopes in Table 5.3 that require some consideration, *AGY* serine-A4, *AGR* arginine-A4 and *UAY* tyrosine-A4. The *AGR* codons are used as termination signals in vertebrate mitochondrial genomes and consequently, the frequency of these codons is zero for many of the data points, regardless of the mutational pressures exhibited at fourfold degenerate sites. Even when vertebrate taxa are removed from the analysis, *AGR* arginine codons fail to show a strong positive correlation with fourfold degenerate site composition. Several additional amino acids, serine (*TCN*), aspartic acid (*GAY*), glutamic acid (*GAR*), isoleucine (*ATY*) and methionine (*ATR*) exhibit weak negative correlations with one of two bases found at the first or second codon position. In all these cases, there is a negative correlation between the fourfold degenerate site frequency of the first codon position base and the second codon position base (Figure 5.5). For example, although the frequency of *TCN* serine codons is expected to increase when mutational pressures favor *T* and *C*, these two mutational pressures are negatively correlated in mitochondrial genomes. Consequently, increases in the rate of change from *UNN* codons to *UCN* codons are offset by decreases in rate of change from *UCN* to *NCN* codons.

8. There is no obvious relationship between degree of neutrality estimated by the slope of these regressions and relative mutability of an amino acid estimated from transition probability matrices for nuclear transmembrane (Jones et al. 1994) or vertebrate mitochondrial proteins (Adachi and Hasagawa 1996). These transition probability matrices

are constructed by tallying the co-occurrence of amino acids at individual sites of homologous protein sequence alignments. Relative mutability is a simple function of the diagonal entries in the 20x20 transition matrix (Jones et al. 1994). Note that relative mutability is a somewhat misleading term to describe these estimates, since they are dependent on both mutation and selection pressures. The expectation that amino acids that show a larger slope, or degree of neutrality, in this analysis will have a higher relative mutability conflicts with the observation that several of the amino acids with the largest slopes, including phenylalanine, proline and glycine exhibit very low relative mutability.

Discussion and Conclusions

Fourfold degenerate sites are a reasonable estimator of mutation pressures although there is some compositional variation among fourfold degenerate codon families, presumably due to either translational level selection, dinucleotide mutational biases, or both. Notably, the pattern of variation among codon families is conserved across diverse metazoan genomes. Whether this conservation is more likely to result from selection or mutation biases is debatable.

In mtDNA, where iso-accepting tRNA abundance is not a factor (Asakawa et al. 1991), selection among synonymous codons to optimize translational accuracy or efficiency is likely to arise from codon-anticodon energetics variation (Chapter III). Studies of nuclear and prokaryotic tRNAs suggest that codon-anticodon interaction dynamics are influenced by the primary sequence of the anticodon (Grosjean and Fiers 1982) plus other features of the tRNA including the size of the variable loop (Curran et al. 1995), modified nucleotides adjacent to the anticodon (Houssier and Grosjean 1985), and base stacking interactions (Grosjean et al. 1978). The fourfold degenerate codons and corresponding anticodons are conserved among these mitochondrial genomes. However, homologous metazoan mitochondrial tRNA sequences are approximately 100 times as variable as their nuclear

counterparts (Kumazawa and Nishida 1993). Many mitochondrial tRNAs vary from the highly conserved nuclear tRNA secondary structure (Wolstenholme 1992). However, even the most variant metazoan mitochondrial tRNAs maintain conserved tertiary interactions (Watanabe et al. 1994). In the absence of direct experimental evidence on mitochondrial tRNA binding and its effects on translation accuracy and efficiency, it is not possible to predict the expected pattern of synonymous codon usage variation among fourfold degenerate families from a single genome. Such data would have to be collected for variant homologous tRNAs from many taxa before it would be possible to predict whether or not the patterns of synonymous codon usage should be consistent across taxa.

Alternatively, the universal patterns of variation among codon families could result from conservation of dinucleotide mutational biases. Under this model, the mutational spectrum at the third codon position of a codon family is dependent on the nucleotide at the second codon position and/or the composition of first positions of the following codons. The classic example of a dinucleotide bias against *CpG* arising from methylation induced mutations appears to be unimportant in mtDNA (Tanaka and Ozawa 1994). However, the lack of support for one particular mutational pathway does not exclude all other dinucleotide biases. Tanaka and Ozawa (1994) favor dinucleotide biases, at either the polymerase misincorporation level or in repair efficiency, as an explanation for differences in mutational frequencies among fourfold degenerate codon families from 43 human mitochondrial genomes. A serious limitation for this model is the inability to explain why dinucleotide mutational biases would be conserved across these diverse metazoan taxa while directional pressures on single nucleotides are quite variable. Although the source of variation among fourfold degenerate sites remains obscure, no codon family shows a radical departure from the mean fourfold degenerate site composition.

There are correlations among the relative frequencies of the four bases at fourfold degenerate sites. Under the single *GC* vs. *AT* mutation pressure model, the equilibrium

GC content of neutral sites is a simple function of two rates of mutation. When the mutation pressure acting on each of the four bases individually is considered, the equilibrium frequency of any given base is a function of not only the rates of mutation, but also the relative frequencies of the other three bases at equilibrium. The observed correlations among mutation pressures could arise in any number of ways. The simplest explanation for a correlation between the frequency of two bases at equilibrium is a direct mutational relationship. That is, mutation pressures on *C* and *T* show a negative correlation because in some lineages *C* → *T* transitions predominate and in other lineages *T* → *C* transitions predominate. Note that it is not necessary for there to be an absolute difference in the rate of *T* ↔ *C* transitions. A similar argument can be made for the negative correlation between *G* and *A*. Transitions have obviously played a major role in mitochondrial evolution, as they outnumber transversions in comparisons of recently diverged sequences and saturate faster among increasingly divergent taxa. The specific mutational pathways associated with shifts in the relative frequencies of different types of transitions is unknown. Spontaneous deamination and oxidative damage by free radicals may be especially important contributors to the mitochondrial mutation spectrum. Variation in rates of replication between genomes and the proportion of time any region of a genome remains single stranded during replication could alter rates of deamination. Oxidative damage might be proportional to metabolic rates. If rates of transition are the principle determinant of base composition, the negative correlation between *A* and *T* is indicative of a mutational spectrum that never simultaneously favors *A* → *G* and *C* → *T* transitions on the same strand. This means it never favors *A* → *G* transitions on one strand and *G* → *A* transitions on the other strand. That the difference in the mutational spectra of the two strands is not radical is appealingly consistent with the idea that both strands are replicated by the same polymerase and should experience the same mutagenic environment for at least some portion of the replication cycle. Regardless of the reason they arise, negative

correlations appear to have a balancing effect on the frequency of amino acids encoded by the two bases with opposing mutational pressures.

First and second codon positions each show positive correlations with all four measures of mutation pressure. The base composition of these sites reflects a balance between mutation pressures and selection for functional mitochondrial proteins. The metazoan mitochondrial genome encodes 12-13 proteins including subunits of all three major respiratory enzyme complexes and ATP synthetase. The high frequency of hydrophobic amino acid residues reflects the fact that these complexes are embedded in the inner mitochondrial membrane. Base substitutions at the second codon position are likely to lead to nonconservative amino acid replacements. For example, a second position substitution from any base to A, replaces a hydrophobic residue with either aspartic or glutamic acid. Thus, selection is likely to constrain the composition of second codon positions (Naylor et al. 1995) to maintain an appropriately charged protein. The genetic code is degenerate at first codon positions for leucine (*TTR* and *CTN*) in all mitochondrial genomes and also arginine (*CGN* and *AGR*) in some genomes. Both the hydrophobicity constraint and the difference in degeneracy are likely to contribute to the difference in the response of first and second codon positions to mutation pressures. All 13 mitochondrial genes individually show positive correlations between first and second position base composition and mutation pressures. Thus no gene product is under such strong selection for amino acid sequence that it is exempt from the effects of DNA level processes on protein composition. However, non-homologous genes show consistent differences in response to mutation pressures that are likely to reflect variation in overall selection acting on the individual genes.

Some amino acids show a greater response to mutation pressures than other amino acids. Jukes and Bhushan (1986) observed an increase in phenylalanine, asparagine and tyrosine and a reduction in alanine and proline in the *AT* rich *Drosophila* genome, relative

to the frequency of these amino acids in five vertebrate genomes. This analysis of strand-specific mutation pressures shows consistent positive correlations of phenylalanine (*TTY*) with *T4*, asparagine (*AA_Y*) with *A4*, tyrosine (*TAY*) with *T4*, alanine (*GCN*) with *C4*, and proline (*CCN*) with *C4*. However, this study also shows that tyrosine is not positively correlated with *A4*, and alanine is not positively correlated with *G4*. Therefore, the distinction made between the protein composition of *AT* rich versus *GC* rich genomes is truly a distinction between *A* rich genomes and *C* rich genomes for these two amino acids. Jukes and Bhushan (1986) also observed that there were no clear trends in the usage of isoleucine, lysine, methionine, glycine or arginine. This analysis shows that isoleucine-*A4*, methionine-*A4*, and especially glycine-*G4* do show a significant positive relationship. In addition, consideration of mutation pressures acting on each base individually reveal strong correlations for valine and threonine with fourfold degenerate site base composition, and a number of weaker correlations for amino acids that were not even considered by Jukes and Bhushan (1986) because a *GC* vs *AT* mutational pressure makes no prediction about usage of amino acids encoded by one *GC* base pair and one *AT* base pair at the first and second codon positions.

The obvious consequence of variation in mutation pressures among lineages and response of amino acid sequence to these biases, is variation in the pattern of amino acid replacement. Just as base compositional variation is indicative of a nonstationary substitution matrix, amino acid compositional variation is indicative of a nonstationary replacement process. Like DNA sequence based phylogenies, protein phylogenies must be constructed assuming an implicit or explicit model of evolution. Amino acid transition probability matrices form the basis of these models. Recently, Adachi and Hasagawa (1996) used maximum likelihood methods to estimate a transition probability matrix specific to mitochondrial proteins and illustrated how this matrix differs from estimates based on overall nuclear proteins or even nuclear-encoded transmembrane proteins. This

analysis of nucleotide and amino acid composition suggests that a single mitochondrial transition matrix may still not truly represent the patterns of amino acid replacement. Protein sequences are most useful for deep-branch phylogenetics where nucleotide sequences have experienced multiple substitutions at individual sites, obscuring the extent of divergence among distantly related taxa. Thus, if mutational pressures differ most among widely divergent taxa, they may have a confounding effect on the data that appear best suited for resolving these level relationships.

In summary, the amino acid composition of mitochondrial proteins may be strongly affected by mutational biases and the strand-specific nature of mitochondrial directional mutation pressures necessitates consideration of each base individually. The balance between mutational pressures and selection is complex. The overall response of the protein composition to mutation pressure is a sum of the differential response of first and second codon positions. The response varies among non-homologous genes and among individual amino acids. Variation in mutation pressures among taxa may have implications for phylogeny reconstruction based on protein sequences.

LITERATURE CITED

- Adachi, J. and M. Hasegawa. (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* 42:459-468.
- Akashi, H. (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* 139:1067-1076.
- Anderson, S., A.T. Bankier, B.G. Barrell, M.H.L. de Bruijn, A.R. Coulson, J. Drouin, I.C. Eperon, D.P. Nierlich, B.A. Roe, F. Sanger, P.H. Schreier, A.J.H. Smith, R. Staden and I.G. Young (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290:457-465.
- Anderson, S., M.H.L. de Bruijn, A.R. Coulson, I.C. Eperon, F. Sanger and I.G. Young. (1982) Complete sequence of bovine mitochondrial DNA: Conserved features of the mammalian mitochondrial genome. *J. Mol. Biol.* 156:683-717.
- Arnason, U., A. Gullberg and B. Widegren. (1991) The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus*. *J. Mol. Evol.* 33:556-568.
- Arnason, U. and A. Gullberg. (1993) Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. *J. Mol. Evol.* 37:312-322.
- Arnason, U., A. Gullberg, E. Johnsson and C. Ledje. (1993) The nucleotide sequence of the mitochondrial DNA molecule of the grey seal *Halichoerus grypus* and a comparison with mitochondrial sequences of other true seals. *J. Mol. Evol.* 37:323-330.
- Arnason, U. and E. Johnsson. (1992) The complete mitochondrial DNA sequence of the harbor seal, *Phoca vitulina*. *J. Mol. Evol.* 34:493-505.
- Asakawa, S., Y. Kumazawa, T. Araki, H. Himeno, K. Miura and K. Watanabe. (1991) Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. *J. Mol. Evol.* 32:511-520.
- Attardi, G., P. Cantatore, A. Chomyn, S. Crews, R. Gelfand, C. Merkel, J. Montoya and D. Ojala. (1982) A comprehensive view of mitochondrial gene expression in human cells. In: P. Slonimski, P. Borst and G. Attardi (eds.), *Mitochondrial Genes*. Cold Spring Harbor Laboratory, pp 51-71.
- Ballard, J.W.O. and M. Kreitman. (1994) Unraveling selection in the mitochondrial genome of *Drosophila*. *Genetics* 138:757-772.

- Beard, C.B., D.M. Hamm and F.H. Collins. (1993) The mitochondrial genome of the mosquito *Anopheles gambiae*: DNA sequence, genome organization, and comparisons with mitochondrial sequences of other insects. *Insect Mol. Biol.* 2:103-104.
- Bibb, M.J., R.A. Van Etten, C.T. Wright, M.W. Walberg and D.A. Clayton. (1981) Sequence and gene organization of mouse mitochondrial DNA. *Cell* 26:167-180.
- Bogenhagen, D. and D.A. Clayton. (1977) Mouse L cell mitochondrial DNA molecules are selected randomly for replication throughout the cell cycle. *Cell* 11:719-727.
- Boore, J.L. and W.M. Brown. (1995) Complete sequence of the mitochondrial DNA of the annelid worm *Lumbricus terrestris*. *Genetics* 141:305-319.
- Boore, J.L. and W.M. Brown. (1994) Complete DNA sequence of the mitochondrial genome of the black chiton, *Katharina tunicata*. *Genetics* 138:423-443.
- Brown, G.G. and M.V. Simpson. (1982) Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. *Proc. Natl. Acad. Sci. USA* 79:3246-3250.
- Brown, W.M. (1981) Mechanisms of evolution in animal mitochondrial DNA. *Annals NY Acad. Sci.* 361:119-134.
- Brown, W.M., E.M. Prager, A. Wang, and A.C. Wilson. (1982) Mitochondrial DNA sequences of primates: Tempo and mode of evolution. *J. Mol. Evol.* 18:225-239.
- Bulmer, M. (1985) Neighboring base effects on substitution rates in pseudogenes. *Mol. Biol. Evol.* 3:322-329.
- Bulmer, M. (1990) The effect of context on synonymous codon usage in genes with low codon bias. *Nucl. Acids Res.* 18:2869-2873.
- Bulmer, M. (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897-907.
- Cantatore, P., M. Roberti, G. Rainaldi, M.N. Gadaleta and C. Saccone. (1989) The complete nucleotide sequence, gene organization, and genetic code of the mitochondrial genome of *Paracentrotus lividus*. *J. Biol. Chem.* 264(19):10965-10975.
- Caswell, H. (1989) *Matrix Population Models*. Sinauer Associates, Inc., Sunderland, Massachusetts, pp 34-39.
- Chang, D.D. and D.A. Clayton. (1985) Priming of human mitochondrial DNA replication occurs at the light-strand promoter. *Proc. Natl. Acad. Sci. USA* 82:351-355.
- Chang, Y.S., F.L. Huang and T. Lo. (1994) The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. *J. Mol. Evol.* 38:138-155.

- Chargaff, E. (1950) Chemical specificity of the nucleic acids and mechanisms of their enzymatic degradation. *Experimentia* 6:201-240.
- Christensen, R. (1990) *Log-Linear Models*. Springer-Verlag, New York.
- Clary, D. and D. Wolstenholme. (1985) The mitochondrial molecule of *Drosophila yakuba*: nucleotide sequence, gene organization, and genetic code. *J. Mol. Evol.* 22:252-271.
- Clayton, D.A. (1982) Replication of animal mitochondrial DNA. *Cell* 28:693-705.
- Clayton, D.A. (1992) Transcription and replication of animal mitochondrial DNAs. *Int. Rev. Cytol.* 141:217-232.
- Collins, T.M, P.H. Wimberger and G.P. Naylor. (1994) Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.* 43:482-496.
- Crozier, R.H. and Y.C. Crozier. (1993) The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics* 133:97-117.
- Curran, J.F., E.S. Poole, W.P. Tate and B.L. Gross. (1995) Selection of aminoacyl-tRNAs at sense codons: the size of the tRNA variable loop determines whether the immediate 3' nucleotide to the codon has a context effect. *Nucleic Acids Res.* 23:4104-4108.
- De Giorgi, C., F. De Luca and C. Saccone. (1991a) Mitochondrial DNA in the sea urchin *Arbacia lixula*: nucleotide sequence differences between two polymorphic molecules indicate asymmetry of mutations. *Gene* 103:249-252.
- De Giorgi, C., C. Lanave, M. Musci and C. Saccone. (1991b) Mitochondrial DNA in the sea urchin *Arbacia lixula*: evolutionary inferences from nucleotide sequence analysis. *Mol. Biol. Evol.* 8:515-529.
- Delorme, M.-O. and A. Henaut. (1991) Codon usage is imposed by the gene location in the transcription unit. *Curr. Genet.* 20:353-358.
- Desjardins, P. and R. Morais. (1990) Sequence and gene organization of the chicken mitochondrial genome: a novel gene order in higher vertebrates. *J. Mol. Biol.* 212:599-634.
- Devereux, J., P. Haeberli and O. Smithies. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nuc. Acids Res.* 12:387-395.
- Eyre-Walker, A.C. (1991) An analysis of codon usage in mammals: selection or mutation bias? *J. Mol. Evol.* 33:442-449.
- Feinberg, S.E. (1980) *The Analysis of Cross Classified Categorical Data*. The MIT Press, Cambridge, Massachusetts.

- Gadaleta, G., G. Pepe, G. De Candia, C. Quagliariello, E. Sbisa and C. Saccone. (1989) The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates. *J. Mol. Evol.* 28:497-516.
- Grosjean, H. and W. Fiers. (1982) Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* 18:199-209.
- Grosjean, H.J., S. de Henau and D.M. Crothers. (1978) On the physical basis for ambiguity in genetic coding interactions. *Proc. Natl. Acad. Sci. USA* 75:610-614.
- Hatzoglou, E., G.C. Rodakis and R. Lecanidou. (1995) The complete sequence of the mitochondrial genome of the land snail *Albinaria coerulea*. *Genetics* 140:1353-1366.
- Himeno, H., H. Masaki, T. Kawai, T. Ohta, I. Kumagai, K. Miura and K. Watanabe. (1987) Unusual genetic codes and a novel gene structure for Ser-tRNA-AGY in starfish mitochondrial DNA. *Gene* 56:219-230.
- Hoffmann, R.J., J.L. Boore and W.M. Brown. (1992) A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. *Genetics* 131:397-412.
- Horai, S., K. Hayasaka, R. Kondo, K. Tsugane and N. Takahata. (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. U.S.A.* 92:532-536.
- Houssier, C. and H. Grosjean. (1985) Temperature jump relaxation studies on the interactions between transfer RNAs with complementary anticodons. The effect of modified bases adjacent to the anticodon triplet. *J. Biomol. Struct. Dyn.* 3:387-408.
- Ikemura, T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2:13-34.
- Irwin, D.M., T.D. Kocher and A.C. Wilson. (1991) Evolution of the cytochrome b gene of mammals. *J. Mol. Evol.* 32:128-144.
- Jacobs, H.T., D.J. Elliott, V.B. Math and A. Farquharson. (1988) Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. *J. Mol. Biol.* 202:185-217.
- Janke, A., G. Feldmaier-Fuchs, W.K. Thomas, A. von Haeseler and S. Paabo. (1994) The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics* 137:243-256.
- Jermiin, L.S., D. Graur, R.M. Lowe and R.H. Crozier. (1994) Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. *J. Mol. Evol.* 39:160-173.

- Jones, D.T., W.R. Taylor and J.M. Thornton. (1994) A mutation data matrix for transmembrane proteins. *FEBS Letters* 339:269-275.
- Jukes, T.H. and C.R. Cantor. (1969) Evolution of protein molecules. In: H.N. Munro (ed.) *Mammalian protein metabolism*. Academic Press, New York, pp 21-132.
- Jukes, T.H. and V. Bhushan. (1986) Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. *J. Mol. Evol.* 24:39-44.
- Kimura, M. (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111-120.
- Kimura, M. (1983) *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, England.
- Knight, A. and D.P. Mindell. (1993) Substitution bias, weighting of DNA sequence evolution, and the phylogenetic position of Fea's viper. *Syst. Biol.* 42:18-31.
- Kocher, T.D. and A.C. Wilson. (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding region. In: Osawa S, T Honjo (eds), *Evolution of life: fossils, molecules and culture*. Springer-Verlag, Tokyo, pp 391-413.
- Kondo, R., S. Horai, Y. Satta and N. Takahata. (1993) Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. *J. Mol. Evol.* 36:517-531.
- Kumazawa, Y. and M. Nishida. (1993) Sequence evolution of mitochondrial tRNA genes and deep-branch animal phylogenetics. *J. Mol. Evol.* 37:380-398.
- Kunkel, T.A. (1985) The mutational specificity of DNA polymerases-alpha and -gamma during *in vitro* DNA synthesis. *J. Biol. Chem.* 260:12866-12874.
- Lanave, C., G. Preparata, C. Saccone and G. Serio. (1984) A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* 20:86-93.
- Lee, W.J. and T.D. Kocher. (1995) Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome: early establishment of the vertebrate genome structure. *Genetics* 139:873-887.
- Lindahl, T. (1993) Instability and decay of the primary structure of DNA. *Nature* 362:709-715.
- Lloyd, A.T. and P.M. Sharp (1992) Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. *Nucl. Acids Res.* 20:5289-5295.
- Margulis, L. (1970) *Origin of eukaryotic cells*. Yale University Press. New Haven, CT.

- Merriwether, D.A., A.G. Clark, S.W. Ballinger, T.G. Schurr, H. Soodyall, T. Jenkins, S.T. Sherry and D.C. Wallace. (1991) The structure of human mitochondrial DNA variation. *J. Mol. Evol.* 33:543-555.
- Naylor, G.J.P., T.M. Collins and W.M. Brown. (1995) Hydrophobicity and phylogeny. *Nature* 373:565-566.
- Okimoto, R., J.L. Macfarlane, D.O. Clary and D.R. Wolstenholme. (1992) The mitochondrial genomes of two nematodes, *Caenorhabditis elegans* and *Ascaris suum*. *Genetics* 130:471-498.
- Palumbi, S.R. and B.D. Kessing. (1991) Population Biology of the trans-arctic exchange: mtDNA sequence similarity between pacific and atlantic sea urchins. *Evol.* 45:1790-1805.
- Perez, M.L., J.R. Valverde, B. Batuecas, F. Amat, R. Marco and R. Garesse. (1994) Speciation in the *Artemia* genus: mitochondrial DNA analysis of bisexual and parthenogenetic brine shrimps. *J. Mol. Evol.* 38:156-168.
- Perna, N.T. and T.D. Kocher. (1995a) Unequal base frequencies and the estimation of substitution rates. *Mol. Biol. Evol.* 12:359-361.
- Perna, N.T. and T.D. Kocher. (1995b) Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J. Mol. Evol.* 41:353-358.
- Rand, D.M., M. Dorfsman, and L.M. Kann. (1994) Neutral and non-neutral evolution of *Drosophila* mitochondrial DNA. *Genetics* 138:741-756.
- Rzhetsky, A. and M. Nei. (1995) Tests of the applicability of several substitution models for DNA sequence data. *Mol. Biol. Evol.* 12:131-151.
- Saccone, C., C. Lanave, G. Pesole and E. Sbisà. (1993) Peculiar features and evolution of mitochondrial genome in mammals. In: Di Mauro and Wallace (eds.) *Mitochondrial DNA in human pathology*. Raven Press, New York, pp 27-39.
- Sharp, P.M. and A.T. Lloyd. (1993) Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. *Nucl. Acids Res.* 21:179-183.
- Shields, D.C. and P.M. Sharp. (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. *Nucl. Acids Res.* 15:8023-8041.
- Sidow A. and W.K. Thomas. (1994) A molecular evolutionary framework for eukaryotic model organisms. *Current Biology* 4(7):596-603.
- Stenico, M., A.T. Lloyd and P.M. Sharp. (1994) Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases. *Nucl. Acids Res.* 22:2437-2446.

- Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl. Acad. Sci. USA* 48:582-592.
- Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci. USA* 85:2653-2657.
- Sueoka, N. (1992) Directional mutation pressure, selective constraints, and genetic equilibria. *J. Mol. Evol.* 34:95-114.
- Sueoka, N. (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.* 40:318-325.
- Tamura, K. (1992) The rate and pattern of nucleotide substitution in *Drosophila* mitochondrial DNA. *Mol. Biol. Evol.* 9:814-825.
- Tamura, K. and M. Nei. (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* 10:512-526.
- Tanaka, M. and T. Ozawa. (1994) Strand asymmetry in human mitochondrial DNA mutations. *Genomics* 22:327-335.
- Tapper, D.P. and D.A. Clayton. (1981) Mechanism of replication of human mitochondrial DNA. *J. Biol. Chem.* 256:5109-5115.
- Terrett, J.A., S. Miles and R.H. Thomas. Complete DNA sequence of the mitochondrial genome of *Cepaea nemoralis*. *J. Mol. Evol.*, in press.
- Thomas, W.K. and A.C. Wilson. (1991) Evolution by base substitution in animal mitochondrial DNA. (unpublished manuscript).
- Tzeng, C.-S., C.-F. Hui, S.-C. Shen and P.C. Huang. (1992) The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome: conservation and variation among vertebrates. *Nucl. Acids Res.* 20:4853-4858.
- Vigilant, L., R. Pennington, H. Harpending, T.D. Kocher and A.C. Wilson. (1991) Mitochondrial DNA sequences in single hairs from a southern African population. *Proc. Natl. Acad. Sci. USA* 86:9350-9354.
- Watanabe, Y., H. Tsurui, T. Ueda, R. Furushima, S. Takamiya, K. Kita, K. Nishikawa and K. Watanabe. (1994) Primary and higher order structures of nematode (*Ascaris suum*) mitochondrial tRNAs lacking either the T or D stem. *J. Biol. Chem.* 269:22902-22906.
- Wolfe, K.H., P.M. Sharp and W.-H. Li. (1989) Mutation rates differ among regions of the mammalian genome. *Nature* 337:283-285.
- Wolstenholme, D. (1992) Animal mitochondrial DNA: Structure and evolution. *Int. Rev. Cytol.* 141:173-216.

- Xiong B. and T.D. Kocher. (1993) Phylogeny of sibling species of *Simulium venustum* and *S. verecundum* (Diptera:Simuliidae) based on sequences of the mitochondrial 16S rRNA gene. *Mol. Phyl. Evol.* 2: 293-303.
- Xu, X. and U. Arnason. (1994) The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene* 148:357-362.
- Yokobori, S. and S. Paabo. (1995) tRNA editing in metazoans. *Nature* 377:490.
- Zardoya, R., J.M. Bautista and A. Garrido-Pertierra. (1995) The complete nucleotide sequence of the mitochondrial DNA of the Rainbow trout, *Onchorhynchus mykiss*. *J. Mol. Evol.* 41:942-951.

APPENDICES

Appendix A: PASCAL Source Code for USAGE

```
program USAGE (input, output, sequence, results, range);

type txt=text;
   chararray=packed array [1..3] of char;

var sequence:txt;
    comp: txt;
    codon:txt;
    seqfilename, rangefilename: string;
    range:txt;
    gene:chararray;
    strand: integer;
    start: integer;
    stop: integer;

procedure countforward(var sequence: txt; var comp:txt; var codon:txt;
    var genename:chararray; var strandnum, first, last:integer);

var ch, ch1, ch2, ch3:char;
    genelength, i, j, k:integer;
    CTA,CTC,CTG,CTT:integer;    {leu}
    GTA,GTC,GTG,GTT:integer;    {val}
    TCA,TCC,TCG,TCT:integer;    {ser}
    CCA,CCC,CCG,CCT:integer;    {pro}
    ACA,ACC,ACG,ACT:integer;    {thr}
    GCA,GCC,GCG,GCT:integer;    {ala}
    CGA,CGC,CGG,CGT:integer;    {arg}
    GGA,GGC,GGG,GGT:integer;    {gly}
    TTT,TTC,TTA,TTG:integer;    {phe/leu}
    ATT,ATC,ATA,ATG:integer;    {ile/met}
    TAT,TAC,TAA,TAG:integer;    {tyr/stop}
    CAT,CAC,CAA,CAG:integer;    {his/gln}
    AAT,AAC,AAA,AAG:integer;    {asn/lys}
    GAT,GAC,GAA,GAG:integer;    {asp/glu}
    TGT,TGC,TGA,TGG:integer;    {cys/stop/trp}
    AGT,AGC,AGA,AGG:integer;    {ser/arg}
    G1,A1,T1,C1,G2,A2,T2,C2,G3,A3,T3,C3,G4,A4,T4,C4:integer;
    FG1,FA1,FT1,FC1,FG2,FA2,FT2,FC2,FG3,FA3,FT3,FC3,FG4,FA4,FT4,
    FC4:real;
    tot1,tot2,tot3,tot4:real;
    phe,leu2,leu4,ile,met,val,ser,pro,thr,ala,tyr,his:real;
    gln,asn,lys,asp,glu,cys,trp,arg4,ser2,arg2,gly:real;
```

begin

CTA:=0;
CTC:=0;
CTG:=0;
CTT:=0;

GTA:=0;
GTC:=0;
GTG:=0;
GTT:=0;

TCA:=0;
TCC:=0;
TCG:=0;
TCT:=0;

CCA:=0;
CCC:=0;
CCG:=0;
CCT:=0;

ACA:=0;
ACC:=0;
ACG:=0;
ACT:=0;

GCA:=0;
GCC:=0;
GCG:=0;
GCT:=0;

CGA:=0;
CGC:=0;
CGG:=0;
CGT:=0;

GGA:=0;
GGC:=0;
GGG:=0;
GGT:=0;

TTA:=0;
TTC:=0;
TTG:=0;
TTT:=0;

ATA:=0;
ATC:=0;
ATG:=0;
ATT:=0;

TAA:=0;
TAC:=0;
TAG:=0;
TAT:=0;

CAA:=0;
CAC:=0;
CAG:=0;
CAT:=0;

AAA:=0;
AAC:=0;
AAG:=0;
AAT:=0;

GAA:=0;
GAC:=0;
GAG:=0;
GAT:=0;

TGA:=0;
TGC:=0;
TGG:=0;
TGT:=0;

AGA:=0;
AGC:=0;
AGG:=0;
AGT:=0;

G1:=0;
A1:=0;
T1:=0;
C1:=0;

G2:=0;
A2:=0;
T2:=0;
C2:=0;

G3:=0;
A3:=0;
T3:=0;
C3:=0;

G4:=0;
A4:=0;
T4:=0;
C4:=0;

fG1:=0;
fA1:=0;
fT1:=0;
fC1:=0;

fG2:=0;
fA2:=0;
fT2:=0;
fC2:=0;

```

fg3:=0;
fa3:=0;
ft3:=0;
fc3:=0;

fg4:=0;
fa4:=0;
ft4:=0;
fc4:=0;

tot1:=0;
tot2:=0;
tot3:=0;
tot4:=0;

phe:=0;
leu2:=0;
leu4:=0;
ile:=0;
met:=0;
val:=0;
ser:=0;
pro:=0;
thr:=0;
ala:=0;
tyr:=0;
his:=0;
gln:=0;
asn:=0;
lys:=0;
asp:=0;
glu:=0;
cys:=0;
trp:=0;
arg4:=0;
ser2:=0;
arg2:=0;
gly:=0;

reset(sequence);
first:=first-1;
i:=0;
while i<first do           {find the start of gene}
  begin
    read(sequence, ch);
    i:=i+1;
    if eoln(sequence) then i:=i-1;
  end;

j:=first;
repeat
  read(sequence, ch1);
  if ch1=' ' then read(sequence, ch1);
  read(sequence, ch2);
  if ch2=' ' then read(sequence, ch2);
  read(sequence, ch3);

```

```

        if ch3=' ' then read(sequence, ch3);
ch1:=upcase(ch1);
ch2:=upcase(ch2);
ch3:=upcase(ch3);

if j=first then write(ch1, ch2, ch3,'.....');
if j+6>last then write(ch1, ch2 ,ch3);

if (ch1='C') and (ch2='T') then      {leu}
    case ch3 of
        'A':CTA:=CTA+1;
        'C':CTC:=CTC+1;
        'G':CTG:=CTG+1;
        'T':CTT:=CTT+1;
    end;

if (ch1='G') and (ch2='T') then      {val}
    case ch3 of
        'A':GTA:=GTA+1;
        'C':GTC:=GTC+1;
        'G':GTG:=GTG+1;
        'T':GTT:=GTT+1;
    end;

if (ch1='T') and (ch2='C') then      {ser}
    case ch3 of
        'A':TCA:=TCA+1;
        'C':TCC:=TCC+1;
        'G':TCG:=TCG+1;
        'T':TCT:=TCT+1;
    end;

if (ch1='C') and (ch2='C') then      {pro}
    case ch3 of
        'A':CCA:=CCA+1;
        'C':CCC:=CCC+1;
        'G':CCG:=CCG+1;
        'T':CCT:=CCT+1;
    end;

if (ch1='A') and (ch2='C') then      {thr}
    case ch3 of
        'A':ACA:=ACA+1;
        'C':ACC:=ACC+1;
        'G':ACG:=ACG+1;
        'T':ACT:=ACT+1;
    end;

if (ch1='G') and (ch2='C') then      {ala}
    case ch3 of
        'A':GCA:=GCA+1;
        'C':GCC:=GCC+1;
        'G':GCG:=GCG+1;
        'T':GCT:=GCT+1;
    end;

```

```

if (ch1='C') and (ch2='G') then      {arg}
  case ch3 of
    'A':CGA:=CGA+1;
    'C':CGC:=CGC+1;
    'G':CGG:=CGG+1;
    'T':CGT:=CGT+1;
  end;

if (ch1='G') and (ch2='G') then      {gly}
  case ch3 of
    'A':GGA:=GGA+1;
    'C':GGC:=GGC+1;
    'G':GGG:=GGG+1;
    'T':GGT:=GGT+1;
  end;

if (ch1='T') and (ch2='T') then      {phe/leu}
  case ch3 of
    'A':TTA:=TTA+1;
    'C':TTC:=TTC+1;
    'G':TTG:=TTG+1;
    'T':TTT:=TTT+1;
  end;

if (ch1='A') and (ch2='T') then      {ile/met}
  case ch3 of
    'A':ATA:=ATA+1;
    'C':ATC:=ATC+1;
    'G':ATG:=ATG+1;
    'T':ATT:=ATT+1;
  end;

if (ch1='T') and (ch2='A') then      {tyr/stop}
  case ch3 of
    'A':TAA:=TAA+1;
    'C':TAC:=TAC+1;
    'G':TAG:=TAG+1;
    'T':TAT:=TAT+1;
  end;

if (ch1='C') and (ch2='A') then      {his/gln}
  case ch3 of
    'A':CAA:=CAA+1;
    'C':CAC:=CAC+1;
    'G':CAG:=CAG+1;
    'T':CAT:=CAT+1;
  end;

if (ch1='A') and (ch2='A') then      {asn/lys}
  case ch3 of
    'A':AAA:=AAA+1;
    'C':AAC:=AAC+1;
    'G':AAG:=AAG+1;
    'T':AAT:=AAT+1;
  end;

```



```

phe:6:3,leu2:6:3,leu4:6:3,ile:6:3,met:6:3,val:6:3,ser:6:3,pro:6:3,
thr:6:3,ala:6:3,tyr:6:3,his:6:3,gln:6:3,asn:6:3,lys:6:3,asp:6:3,
glu:6:3,cys:6:3,trp:6:3,arg4:6:3,ser2:6:3,arg2:6:3,gly:6:3);
writeln;

```

```

writeln(codon, seqfilename, ' ', genename, strandnum, ' F ',
      TTT, ' ', TTC, ' ', TTA, ' ', TTG, ' ',
      CTT, ' ', CTC, ' ', CTA, ' ', CTG, ' ',
      ATT, ' ', ATC, ' ', ATA, ' ', ATG, ' ',
      GTT, ' ', GTC, ' ', GTA, ' ', GTG, ' ',
      TCT, ' ', TCC, ' ', TCA, ' ', TCG, ' ',
      CCT, ' ', CCC, ' ', CCA, ' ', CCG, ' ',
      ACT, ' ', ACC, ' ', ACA, ' ', ACG, ' ',
      GCT, ' ', GCC, ' ', GCA, ' ', GCG, ' ',
      TAT, ' ', TAC, ' ', TAA, ' ', TAG, ' ',
      CAT, ' ', CAC, ' ', CAA, ' ', CAG, ' ',
      AAT, ' ', AAC, ' ', AAA, ' ', AAG, ' ',
      GAT, ' ', GAC, ' ', GAA, ' ', GAG, ' ',
      TGT, ' ', TGC, ' ', TGA, ' ', TGG, ' ',
      CGT, ' ', CGC, ' ', CGA, ' ', CGG, ' ',
      AGT, ' ', AGC, ' ', AGA, ' ', AGG, ' ',
      GGT, ' ', GGC, ' ', GGA, ' ', GGG, ' ');

```

```

writeln(codon, seqfilename, ' ', genename, strandnum, ' P',
      (TTT/(TTT+TTC)):9:3,
      (TTC/(TTT+TTC)):9:3,
      (TTA/(TTA+TTG)):9:3,
      (TTG/(TTA+TTG)):9:3,
      (CTT/(CTT+CTC+CTA+CTG)):9:3,
      (CTC/(CTT+CTC+CTA+CTG)):9:3,
      (CTA/(CTT+CTC+CTA+CTG)):9:3,
      (CTG/(CTT+CTC+CTA+CTG)):9:3,
      (ATT/(ATT+ATC)):9:3,
      (ATC/(ATT+ATC)):9:3,
      (ATA/(ATA+ATG)):9:3,
      (ATG/(ATA+ATG)):9:3,
      (GTT/(GTT+GTC+GTA+GTG)):9:3,
      (GTC/(GTT+GTC+GTA+GTG)):9:3,
      (GTA/(GTT+GTC+GTA+GTG)):9:3,
      (GTG/(GTT+GTC+GTA+GTG)):9:3,
      (TCT/(TCT+TCC+TCA+TCG)):9:3,
      (TCC/(TCT+TCC+TCA+TCG)):9:3,
      (TCA/(TCT+TCC+TCA+TCG)):9:3,
      (TCG/(TCT+TCC+TCA+TCG)):9:3,
      (CCT/(CCT+CCC+CCA+CCG)):9:3,
      (CCC/(CCT+CCC+CCA+CCG)):9:3,
      (CCA/(CCT+CCC+CCA+CCG)):9:3,
      (CCG/(CCT+CCC+CCA+CCG)):9:3,
      (ACT/(ACT+ACC+ACA+ACG)):9:3,
      (ACC/(ACT+ACC+ACA+ACG)):9:3,
      (ACA/(ACT+ACC+ACA+ACG)):9:3,
      (ACG/(ACT+ACC+ACA+ACG)):9:3,
      (GCT/(GCT+GCC+GCA+GCG)):9:3,
      (GCC/(GCT+GCC+GCA+GCG)):9:3,
      (GCA/(GCT+GCC+GCA+GCG)):9:3,
      (GCG/(GCT+GCC+GCA+GCG)):9:3,

```

```

(TAT/(TAT+TAC)):9:3,
(TAC/(TAT+TAC)):9:3,
(TAA/(TAG+TAA)):9:3,
(TAG/(TAG+TAA)):9:3,
(CAT/(CAT+CAC)):9:3,
(CAC/(CAT+CAC)):9:3,
(CAA/(CAA+CAG)):9:3,
(CAG/(CAA+CAG)):9:3,
(AAT/(AAT+AAC)):9:3,
(AAC/(AAT+AAC)):9:3,
(AAA/(AAA+AAG)):9:3,
(AAG/(AAA+AAG)):9:3,
(GAT/(GAT+GAC)):9:3,
(GAC/(GAT+GAC)):9:3,
(GAA/(GAA+GAG)):9:3,
(GAG/(GAA+GAG)):9:3,
(TGT/(TGT+TGC)):9:3,
(TGC/(TGT+TGC)):9:3,
(TGA/(TGA+TGG)):9:3,
(TGG/(TGA+TGG)):9:3,
(CGT/(CGT+CGC+CGA+CGG)):9:3,
(CGC/(CGT+CGC+CGA+CGG)):9:3,
(CGA/(CGT+CGC+CGA+CGG)):9:3,
(CGG/(CGT+CGC+CGA+CGG)):9:3,
(AGT/(AGT+AGC)):9:3,
(AGC/(AGT+AGC)):9:3,
(AGA/(AGA+AGG)):9:3,
(AGG/(AGA+AGG)):9:3,
(GGT/(GGT+GGC+GGA+GGG)):9:3,
(GGC/(GGT+GGC+GGA+GGG)):9:3,
(GGA/(GGT+GGC+GGA+GGG)):9:3,
(GGG/(GGT+GGC+GGA+GGG)):9:3);

```

end;

```

procedure countreverse(var sequence: txt; var comp:txt; var codon:txt;
var genename:chararray; var strandnum, first, last:integer);

```

```

var ch, ch1, ch2, ch3:char;
i, j, k, genelength:integer;
CTA, CTC, CTG, CTT:integer; {leu}
GTA, GTC, GTG, GTT:integer; {val}
TCA, TCC, TCG, TCT:integer; {ser}
CCA, CCC, CCG, CCT:integer; {pro}
ACA, ACC, ACG, ACT:integer; {thr}
GCA, GCC, GCG, GCT:integer; {ala}
CGA, CGC, CGG, CGT:integer; {arg}
GGA, GGC, GGG, GGT:integer; {gly}
TTT, TTC, TTA, TTG:integer; {phe/leu}
ATT, ATC, ATA, ATG:integer; {ile/met}
TAT, TAC, TAA, TAG:integer; {tyr/stop}
CAT, CAC, CAA, CAG:integer; {his/gln}
AAT, AAC, AAA, AAG:integer; {asn/lys}
GAT, GAC, GAA, GAG:integer; {asp/glu}
TGT, TGC, TGA, TGG:integer; {cys/stop/trp}

```

```

AGT,AGC,AGA,AGG:integer;      {ser/arg}
G1,A1,T1,C1,G2,A2,T2,C2,G3,A3,T3,C3,G4,A4,T4,C4:integer;
fg1,fa1,ft1,fc1,fg2,fa2,ft2,fc2,fg3,fa3,ft3,fc3,fg4,fa4,ft4,
fc4:real;
tot1,tot2,tot3,tot4:real;
phe,leu2,leu4,ile,met,val,ser,pro,thr,ala,tyr,his:real;
gln,asn,lys,asp,glu,cys,trp,arg4,ser2,arg2,gly:real;

begin

CTA:=0;
CTC:=0;
CTG:=0;
CTT:=0;

GTA:=0;
GTC:=0;
GTG:=0;
GTT:=0;

TCA:=0;
TCC:=0;
TCG:=0;
TCT:=0;

CCA:=0;
CCC:=0;
CCG:=0;
CCT:=0;

ACA:=0;
ACC:=0;
ACG:=0;
ACT:=0;

GCA:=0;
GCC:=0;
GCG:=0;
GCT:=0;

CGA:=0;
CGC:=0;
CGG:=0;
CGT:=0;

GGA:=0;
GGC:=0;
GGG:=0;
GGT:=0;

TTA:=0;
TTC:=0;
TTG:=0;
TTT:=0;

ATA:=0;

```

ATC:=0;
ATG:=0;
ATT:=0;

TAA:=0;
TAC:=0;
TAG:=0;
TAT:=0;

CAA:=0;
CAC:=0;
CAG:=0;
CAT:=0;

AAA:=0;
AAC:=0;
AAG:=0;
AAT:=0;

GAA:=0;
GAC:=0;
GAG:=0;
GAT:=0;

TGA:=0;
TGC:=0;
TGG:=0;
TGT:=0;

AGA:=0;
AGC:=0;
AGG:=0;
AGT:=0;

G1:=0;
A1:=0;
T1:=0;
C1:=0;

G2:=0;
A2:=0;
T2:=0;
C2:=0;

G3:=0;
A3:=0;
T3:=0;
C3:=0;

G4:=0;
A4:=0;
T4:=0;
C4:=0;

fG1:=0;
fA1:=0;

```

fT1:=0;
fC1:=0;
fG2:=0;

fA2:=0;
fT2:=0;
fC2:=0;

fG3:=0;
fA3:=0;
fT3:=0;
fC3:=0;

fG4:=0;
fA4:=0;
fT4:=0;
fC4:=0;

tot1:=0;
tot2:=0;
tot3:=0;
tot4:=0;

phe:=0;
leu2:=0;
leu4:=0;
ile:=0;
met:=0;
val:=0;
ser:=0;
pro:=0;
thr:=0;
ala:=0;
tyr:=0;
his:=0;
gln:=0;
asn:=0;
lys:=0;
asp:=0;
glu:=0;
cys:=0;
trp:=0;
arg4:=0;
ser2:=0;
arg2:=0;
gly:=0;

reset(sequence);
first:=first-1;
i:=0;
while i<first do           {find start of gene}
  begin
    read(sequence, ch);
    i:=i+1;
    if eoln(sequence) then i:=i-1;
  end;

```

```

genelength:=last-first;
if (genelength MOD 3)>0 then
  begin
    writeln('**GENE LENGTH NOT A MULTIPLE OF 3**');
    k:=0;
    while (genelength MOD 3) > k do      {eliminate partial codon}
      begin
        read(sequence, ch);
        write(ch);
        k:=k+1;
      end;
    write(' ');
  end;

j:=first+(genelength MOD 3);
while j<last do
  begin

    read(sequence, ch3);
    if ch3=' ' then read(sequence, ch3);
    read(sequence, ch2);
    if ch2=' ' then read(sequence, ch2);
    read(sequence, ch1);
    if ch1=' ' then read(sequence, ch1);
    ch1:=upcase(ch1);
    ch2:=upcase(ch2);
    ch3:=upcase(ch3);

    if j=first+(genelength MOD 3) then write(ch3, ch2,
      ch1, '.....');
    if j+4>last then writeln(ch3, ch2 ,ch1);

    if (ch1='G') and (ch2='A') then      {leu}
      case ch3 of
        'T':CTA:=CTA+1;
        'G':CTC:=CTC+1;
        'C':CTG:=CTG+1;
        'A':CTT:=CTT+1;
      end;

    if (ch1='C') and (ch2='A') then      {val}
      case ch3 of
        'T':GTA:=GTA+1;
        'G':GTC:=GTC+1;
        'C':GTG:=GTG+1;
        'A':GTT:=GTT+1;
      end;

    if (ch1='A') and (ch2='G') then      {ser}
      case ch3 of
        'T':TCA:=TCA+1;
        'G':TCC:=TCC+1;
        'C':TCG:=TCG+1;
        'A':TCT:=TCT+1;
      end;
  end;

```

```

if (ch1='G') and (ch2='G') then      {pro}
  case ch3 of
    'T':CCA:=CCA+1;
    'G':CCC:=CCC+1;
    'C':CCG:=CCG+1;
    'A':CCT:=CCT+1;
  end;

if (ch1='T') and (ch2='G') then      {thr}
  case ch3 of
    'T':ACA:=ACA+1;
    'G':ACC:=ACC+1;
    'C':ACG:=ACG+1;
    'A':ACT:=ACT+1;
  end;

if (ch1='C') and (ch2='G') then      {ala}
  case ch3 of
    'T':GCA:=GCA+1;
    'G':GCC:=GCC+1;
    'C':GCG:=GCG+1;
    'A':GCT:=GCT+1;
  end;

if (ch1='G') and (ch2='C') then      {arg}
  case ch3 of
    'T':CGA:=CGA+1;
    'G':CGC:=CGC+1;
    'C':CGG:=CGG+1;
    'A':CGT:=CGT+1;
  end;

if (ch1='C') and (ch2='C') then      {gly}
  case ch3 of
    'T':GGA:=GGA+1;
    'G':GGC:=GGC+1;
    'C':GGG:=GGG+1;
    'A':GGT:=GGT+1;
  end;

if (ch1='A') and (ch2='A') then      {phe/leu}
  case ch3 of
    'T':TTA:=TTA+1;
    'G':TTC:=TTC+1;
    'C':TTG:=TTG+1;
    'A':TTT:=TTT+1;
  end;

if (ch1='T') and (ch2='A') then      {ile/met}
  case ch3 of
    'T':ATA:=ATA+1;
    'G':ATC:=ATC+1;
    'C':ATG:=ATG+1;
    'A':ATT:=ATT+1;
  end;

```



```

if (ch1='A') and (ch2='T') then      {tyr/stop}
  case ch3 of
    'T':TAA:=TAA+1;
    'G':TAC:=TAC+1;
    'C':TAG:=TAG+1;
    'A':TAT:=TAT+1;
  end;

if (ch1='G') and (ch2='T') then      {his/gln}
  case ch3 of
    'T':CAA:=CAA+1;
    'G':CAC:=CAC+1;
    'C':CAG:=CAG+1;
    'A':CAT:=CAT+1;
  end;

if (ch1='T') and (ch2='T') then      {asn/lys}
  case ch3 of
    'T':AAA:=AAA+1;
    'G':AAC:=AAC+1;
    'C':AAG:=AAG+1;
    'A':AAT:=AAT+1;
  end;

if (ch1='C') and (ch2='T') then      {asp/glu}
  case ch3 of
    'T':GAA:=GAA+1;
    'G':GAC:=GAC+1;
    'C':GAG:=GAG+1;
    'A':GAT:=GAT+1;
  end;

if (ch1='A') and (ch2='C') then      {cys/stop/trp}
  case ch3 of
    'T':TGA:=TGA+1;
    'G':TGC:=TGC+1;
    'C':TGG:=TGG+1;
    'A':TGT:=TGT+1;
  end;

if (ch1='T') and (ch2='C') then      {ser/arg}
  case ch3 of
    'T':AGA:=AGA+1;
    'G':AGC:=AGC+1;
    'C':AGG:=AGG+1;
    'A':AGT:=AGT+1;
  end;

j:=j+3;

end;

G1:=GTT+GTC+GTA+GTG+GCT+GCC+GCA+GCG+GAT+GAC+GAA+GAG+GGT+GGC+GGA+GGG;
A1:=ATT+ATC+ATA+ATG+ACT+ACC+ACA+ACG+AAT+AAC+AAA+AAG+AGT+AGC+AGA+AGG;

```



```

trp:=(TGA+TGG)/tot3;
arg4:=(CGT+CGC+CGA+CGG)/tot3;
ser2:=(AGT+AGC)/tot3;
arg2:=(AGA+AGG)/tot3;
gly:=(GGT+GGC+GGA+GGG)/tot3;

writeln;
writeln(comp, seqfilename, ' ', genename, strandnum, ' F ', G1, ' ',
A1, ' ', T1, ' ', C1, ' ', G2, ' ', A2, ' ', T2, ' ', C2, ' ', G3, ' ', A3, ' ',
T3, ' ', C3, ' ', G4, ' ', A4, ' ', T4, ' ', C4, ' ');
writeln(comp, seqfilename, ' ', genename, strandnum, ' P ',
fg1:6:3, fa1:6:3, ft1:6:3, fc1:6:3, fg2:6:3, fa2:6:3, ft2:6:3, fc2:6:3,
fg3:6:3, fa3:6:3, ft3:6:3, fc3:6:3, fg4:6:3, fa4:6:3, ft4:6:3, fc4:6:3,
phe:6:3, leu2:6:3, leu4:6:3, ile:6:3, met:6:3, val:6:3, ser:6:3, pro:6:3,
thr:6:3, ala:6:3, tyr:6:3, his:6:3, gln:6:3, asn:6:3, lys:6:3, asp:6:3,
glu:6:3, cys:6:3, trp:6:3, arg4:6:3, ser2:6:3, arg2:6:3, gly:6:3);

writeln(codon, seqfilename, ' ', genename, strandnum, ' F ',
TTT, ' ', TTC, ' ', TTA, ' ', TTG, ' ',
CTT, ' ', CTC, ' ', CTA, ' ', CTG, ' ',
ATT, ' ', ATC, ' ', ATA, ' ', ATG, ' ',
GTT, ' ', GTC, ' ', GTA, ' ', GTG, ' ',
TCT, ' ', TCC, ' ', TCA, ' ', TCG, ' ',
CCT, ' ', CCC, ' ', CCA, ' ', CCG, ' ',
ACT, ' ', ACC, ' ', ACA, ' ', ACG, ' ',
GCT, ' ', GCC, ' ', GCA, ' ', GCG, ' ',
TAT, ' ', TAC, ' ', TAA, ' ', TAG, ' ',
CAT, ' ', CAC, ' ', CAA, ' ', CAG, ' ',
AAT, ' ', AAC, ' ', AAA, ' ', AAG, ' ',
GAT, ' ', GAC, ' ', GAA, ' ', GAG, ' ',
TGT, ' ', TGC, ' ', TGA, ' ', TGG, ' ',
CGT, ' ', CGC, ' ', CGA, ' ', CGG, ' ',
AGT, ' ', AGC, ' ', AGA, ' ', AGG, ' ',
GGT, ' ', GGC, ' ', GGA, ' ', GGG, ' ');
writeln(codon, seqfilename, ' ', genename, strandnum, ' P ',
(TTT/(TTT+TTC)):9:3,
(TTC/(TTT+TTC)):9:3,
(TTA/(TTA+TTG)):9:3,
(TTG/(TTA+TTG)):9:3,
(CTT/(CTT+CTC+CTA+CTG)):9:3,
(CTC/(CTT+CTC+CTA+CTG)):9:3,
(CTA/(CTT+CTC+CTA+CTG)):9:3,
(CTG/(CTT+CTC+CTA+CTG)):9:3,
(ATT/(ATT+ATC)):9:3,
(ATC/(ATT+ATC)):9:3,
(ATA/(ATA+ATG)):9:3,
(ATG/(ATA+ATG)):9:3,
(GTT/(GTT+GTC+GTA+GTG)):9:3,
(GTC/(GTT+GTC+GTA+GTG)):9:3,
(GTA/(GTT+GTC+GTA+GTG)):9:3,
(GTG/(GTT+GTC+GTA+GTG)):9:3,
(TCT/(TCT+TCC+TCA+TCG)):9:3,
(TCC/(TCT+TCC+TCA+TCG)):9:3,
(TCA/(TCT+TCC+TCA+TCG)):9:3,
(TCG/(TCT+TCC+TCA+TCG)):9:3,
(CCT/(CCT+CCC+CCA+CCG)):9:3,

```

```

(CCC/ (CCT+CCC+CCA+CCG)) :9:3,
(CCA/ (CCT+CCC+CCA+CCG)) :9:3,
(CCG/ (CCT+CCC+CCA+CCG)) :9:3,
(ACT/ (ACT+ACC+ACA+ACG)) :9:3,
(ACC/ (ACT+ACC+ACA+ACG)) :9:3,
(ACA/ (ACT+ACC+ACA+ACG)) :9:3,
(ACG/ (ACT+ACC+ACA+ACG)) :9:3,
(GCT/ (GCT+GCC+GCA+GCG)) :9:3,
(GCC/ (GCT+GCC+GCA+GCG)) :9:3,
(GCA/ (GCT+GCC+GCA+GCG)) :9:3,
(GCG/ (GCT+GCC+GCA+GCG)) :9:3,
(TAT/ (TAT+TAC)) :9:3,
(TAC/ (TAT+TAC)) :9:3,
(TAA/ (TAG+TAA)) :9:3,
(TAG/ (TAG+TAA)) :9:3,
(CAT/ (CAT+CAC)) :9:3,
(CAC/ (CAT+CAC)) :9:3,
(CAA/ (CAA+CAG)) :9:3,
(CAG/ (CAA+CAG)) :9:3,
(AAT/ (AAT+AAC)) :9:3,
(AAC/ (AAT+AAC)) :9:3,
(AAA/ (AAA+AAG)) :9:3,
(AAG/ (AAA+AAG)) :9:3,
(GAT/ (GAT+GAC)) :9:3,
(GAC/ (GAT+GAC)) :9:3,
(GAA/ (GAA+GAG)) :9:3,
(GAG/ (GAA+GAG)) :9:3,
(TGT/ (TGT+TGC)) :9:3,
(TGC/ (TGT+TGC)) :9:3,
(TGA/ (TGA+TGG)) :9:3,
(TGG/ (TGA+TGG)) :9:3,
(CGT/ (CGT+CGC+CGA+CGG)) :9:3,
(CGC/ (CGT+CGC+CGA+CGG)) :9:3,
(CGA/ (CGT+CGC+CGA+CGG)) :9:3,
(CGG/ (CGT+CGC+CGA+CGG)) :9:3,
(AGT/ (AGT+AGC)) :9:3,
(AGC/ (AGT+AGC)) :9:3,
(AGA/ (AGA+AGG)) :9:3,
(AGG/ (AGA+AGG)) :9:3,
(GGT/ (GGT+GGC+GGA+GGG)) :9:3,
(GGC/ (GGT+GGC+GGA+GGG)) :9:3,
(GGA/ (GGT+GGC+GGA+GGG)) :9:3,
(GGG/ (GGT+GGC+GGA+GGG)) :9:3;

end;

{main program}
begin
  writeln('This program generates a table of nucleotide and amino acid
    usage and');
  writeln('a codon usage table. ');
  writeln('Name of sequence file? ');
  readln(seqfilename);
  writeln;
  writeln('Name of file specifying gene boundaries? ');
  readln(rangefilename);

```

```

writeln;
open(sequence, seqfilename);
open(range, rangefilename);
open(comp, 'comp.dat');
open(codon, 'codon.dat');
writeln(comp, 'sequence', ' gen', ' s', ' id ', 'G1 ', 'A1 ', 'T1 ',
        'C1 ', 'G2 ', 'A2 ', 'T2 ', 'C2 ', 'G3 ', 'A3 ', 'T3 ', 'C3 ', 'G4 ',
        'A4 ', 'T4 ', 'C4 ', ' phe ', 'leu2 ', 'leu4 ', 'ile ', 'met ',
        'val ', 'ser ', 'pro ', 'thr ', 'ala ', 'tyr ', 'his ', 'gln ',
        'asn ', 'lys ', 'asp ', 'glu ', 'cys ', 'trp ', 'arg4 ', 'ser2 ',
        'arg2 ', 'gly ');
writeln(codon, 'sequence', ' gen', ' s', ' id ', 'phe ', 'phe ', 'leu ',
        'leu ', 'leu ', 'leu ', 'leu ', 'leu ', 'ile ', 'ile ', 'met ', 'met ',
        'val ', 'val ', 'val ', 'val ', 'ser ', 'ser ', 'ser ', 'ser ', 'pro ',
        'pro ', 'pro ', 'thr ', 'thr ', 'thr ', 'thr ', 'ala ', 'ala ',
        'ala ', 'ala ', 'tyr ', 'tyr ', 'stp ', 'stp ', 'his ', 'his ', 'gln ',
        'gln ', 'asn ', 'asn ', 'lys ', 'lys ', 'asp ', 'asp ', 'glu ', 'glu ',
        'cys ', 'cys ', 'stp ', 'trp ', 'arg ', 'arg ', 'arg ', 'arg ', 'ser ',
        'ser ', 'arg ', 'arg ', 'gly ', 'gly ', 'gly ', 'gly ');
writeln(codon, 'sequence', ' gen', ' s', ' id ',
        'TTT ', 'TTC ', 'TTA ', 'TTG ',
        'CTT ', 'CTC ', 'CTA ', 'CTG ',
        'ATT ', 'ATC ', 'ATA ', 'ATG ',
        'GTT ', 'GTC ', 'GTA ', 'GTG ',
        'TCT ', 'TCC ', 'TCA ', 'TCG ',
        'CCT ', 'CCC ', 'CCA ', 'CCG ',
        'ACT ', 'ACC ', 'ACA ', 'ACG ',
        'GCT ', 'GCC ', 'GCA ', 'GCG ',
        'TAT ', 'TAC ', 'TAA ', 'TAG ',
        'CAT ', 'CAC ', 'CAA ', 'CAG ',
        'AAT ', 'AAC ', 'AAA ', 'AAG ',
        'GAT ', 'GAC ', 'GAA ', 'GAG ',
        'TGT ', 'TGC ', 'TGA ', 'TGG ',
        'CGT ', 'CGC ', 'CGA ', 'CGG ',
        'AGT ', 'AGC ', 'AGA ', 'AGG ',
        'GGT ', 'GGC ', 'GGA ', 'GGG ');

while not EOF(range) do
begin
    readln(range, gene, strand, start, stop);
    writeln(gene, strand, start, stop);
    if strand=1 then countforward(sequence, comp, codon, gene,
        strand, start, stop) else
    if strand=2 then countreverse(sequence, comp, codon, gene,
        strand, start, stop)
    end;
    writeln('The results files are complete. ');
end.

```

Appendix B: Sample Datarange Input File

{This is a sample input file for the honeybee genome. Individual genes can be listed in any order. Column two indicates the polarity of the reading frame using 1 for forward and 2 for reverse complement}

nd2	1	502	1503	
co1	1	1794	3359	
co2	1	3618	4295	
ap8	1	4444	4584	{eliminated ap6 overlap}
ap6	1	4602	5264	{eliminated ap8 overlap}
co3	1	5285	6064	
nd3	1	6185	6538	
nd5	2	6892	8556	
nd4	2	8644	9987	
n4L	2	9991	10254	
nd6	1	10441	10944	
cyb	1	11004	12155	
nd1	2	12302	13219	

Appendix C: Sample Output Log

This program generates a table of nucleotide and amino acid usage and a codon usage table.

Name of sequence file?

apis.dna

Name of file specifying gene boundaries?

apisrange

nd2	1	502	1503
ATC.....TAA			

co1	1	1794	3359
ATA.....TAA			

co2	1	3618	4295
ATT.....TAA			

ap8	1	4444	4584
ATT.....AAA			

ap6	1	4602	5264
ATA.....TAA			

co3	1	5285	6064
ATG.....TAA			

nd3	1	6185	6538
ATA.....TAA			

nd5	2	6892	8556
TTA.....AAT			

nd4	2	8644	9987
TTA.....TAT			

n4L	2	9991	10254
TTA.....AAT			

nd6	1	10441	10944
ATT.....TAA			

cyb	1	11004	12155
ATG.....TAA			

nd1	2	12302	13219
TTA.....AAT			

The results files are complete.